Polskie Towarzystwo Statystyczne
Oddział we Wrocławiu

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Silesian Statistical Review

Nr 7 (13)

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2009

Redaktor Wydawnictwa

Aleksandra Śliwka

Redaktor techniczny

Barbara Łopusiewicz

Korektor

Barbara Cibis

Projekt okładki

Beata Dębska

ADRES REDAKCJI

Katedra Statystyki
Uniwersytetu Ekonomicznego we Wrocławiu
ul. Komandorska 118/120, 53-345 Wrocław
tel. (71) 36-80-356, tel./fax (71) 36-80-357
e-mail: stanislaw.heilpern@ue.wroc.pl

PL ISSN 1644-6739

Spis treści

19. SCIENTIFIC STATISTICAL SEMINAR "MARBURG/KÖLN – WROCŁAW" Burgthann, September 22-25, 2008

EXTENDED ABSTRACTS OF PAPERS

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 7 (13)

PL ISSN 1644-6739

ASYMPTOTIC NORMALITY OF STOPPING TIMES IN SEQUENTIAL CHANGE-POINT ANALYSIS

Josef G. Steinebach (University of Cologne)

1. Introduction

In Gut and Steinebach [2002], we constructed some truncated sequential monitoring procedures for detecting a structural break ("change-point") in a series of counting data, e.g., in the number of claims of an insurance portfolio, which have been observed sequentially at equidistant time-points $t = 0,1, \ldots, n$. Some limiting extreme value asymptotics (as $n \to \infty$) have been derived there under the null hypothesis of "no change", which allow for an asymptotic choice of the critical boundaries in the monitoring schemes such that the false alarm rate can be kept below a prescribed level α. Moreover, some limiting properties under the alternative could also be proved showing that the statistical procedures have asymptotic power 1. The present note reports on the recent work of Gut and Steinebach [2008], which is a continuation of the previous one, the main point being that in [Gut, Steinebach 2008] we look in more detail into the behaviour of the relevant stopping times, in particular the time it takes from the (unknown) change-point until one detects that a change actually has occurred, in other words, asymptotics for stopping times under alternatives are proved.

As in Gut and Steinebach [2002], we observe counting data $N(0), N(1), \ldots, N(n)$ at time-points $t = 0,1, \ldots, n$, where $\{N(t)\}_{0 \le t \le n}$ is a renewal counting process with drift coefficient θ and variance parameter η^2 up to some (unknown) change-point k_n^*, after which it changes to an independent second renewal counting process with drift coefficient θ^* and variance parameter $(\eta^*)^2$. We want to test, e.g., the null hypothesis

$$H_0 : k_n^* = n \quad (\text{"no change"})$$

versus the (two-sided) alternative

$$H_1 : 1 \le k_n^* < n,\ \theta^* \ne \theta \quad (\text{"change at } k_n^*\text{"}),$$

taking sequentially into account the observed counting variables $N(0), N(1), \ldots, N(n)$, where n is a truncation point for our procedure. Our asymptotics below are based on a strong invariance principle (cf. Gut and Steinebach [2002], Proposition 3.1), which shows that, under an r-th moment condition (with some $r > 2$), the counting process $\{N(t)\}_{0 \le t \le n}$ above can almost surely be approximated (with a rate $n^{1/r}$) by a Gaussian process $\{V(t)\}_{0 \le t \le n}$ possessing a similar structure, i.e., also having a drift coefficient θ and variance parameter η^2 up to the change-point k_n^*, and changing thereafter to an independent second Gaussian process with drift coefficient θ^* and variance parameter $(\eta^*)^2$.

In Section 2 we review, for the readers' convenience, some results under the null hypothesis from Gut and Steinebach [2002], before we state our main results on the asymptotic normality of stopping times in Section 3.

2. Stopping times and critical values

From the sequential observations $\{N(k)\}_{k=0,1,\ldots,n}$, we compute the variables

$$Y_k = Y_{k,n} = \frac{N(k) - N(k - h_n) - h_n\theta}{\eta\sqrt{h_n}}, \quad k = h_n, \ldots, n,$$

$$Z_k = Z_{k,n} = \frac{N(k) - k\theta}{\eta\sqrt{k}}, \quad k = k_n, \ldots, n,$$

and the stopping times

$$\tau_n^{(1)} = \min\left\{h_n \le k < n : |Y_k| > c_n^{(1)}\right\},$$

$$\tau_n^{(2)} = \min\left\{k_n \le k < n : |Z_k| > c_n^{(2)}\right\},$$

where $c_n^{(1)}, c_n^{(2)}$ are suitable critical values and h_n, k_n are the lengths of the respective "training periods".

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

Remark 1. For the sake of simplicity, we assume here that the "in-control" parameters θ, η are known, but they can also be replaced by sequential estimates (cf. Gut and Steinebach [2002]).

The critical values $c_n^{(1)}, c_n^{(2)}$ are chosen such that the false alarm rates (asymptotically) attain a prescribed level α, i.e.,

$$P_{H_0}\left(\tau_n^{(1)} < n\right) = P_{H_0}\left(\max_{h_n \le k < n} |Y_k| > c_n^{(1)}\right) \approx \alpha,$$

and

$$P_{H_0}\left(\tau_n^{(2)} < n\right) = P_{H_0}\left(\max_{k_n \le k < n} |Z_k| > c_n^{(2)}\right) \approx \alpha,$$

which can be achieved via the following extreme value asymptotics (see Gut and Steinebach [2002]):

Theorem 1. If $h_n \ll n$, but $\sqrt{h_n} \gg n^{1/r}$, then, under H_0, there are normalizing sequences $\left\{a_n^{(1)}\right\}$ and $\left\{b_n^{(1)}\right\}$ such that

$$a_n^{(1)} \max_{h_n \le k < n} |Y_k| - b_n^{(1)} \to E \quad \text{(two-sided Gumbel)}$$

in distribution (as $n \to \infty$), that is, the critical value $c_n^{(1)}$ can (asymptotically) be chosen as

$$c_n^{(1)} \approx \frac{E_{1-\alpha} + b_n^{(1)}}{a_n^{(1)}} \quad \left(\approx \sqrt{2\log(n/h_n)}\right).$$

Theorem 2. If $k_n \ll n$, but $\sqrt{k_n} \gg n^{1/r}$, then, under H_0, there are normalizing sequences $\left\{a_n^{(2)}\right\}$ and $\left\{b_n^{(2)}\right\}$ such that

$$a_n^{(2)} \max_{k_n \le k < n} |Z_k| - b_n^{(2)} \to E \text{ (two-sided Gumbel)}$$

in distribution (as $n \to \infty$), that is, the critical value $c_n^{(2)}$ can (asymptotically) be chosen as

$$c_n^{(2)} \approx \frac{E_{1-\alpha} + b_n^{(2)}}{a_n^{(2)}} \quad \left(\approx \sqrt{2\log\log(n/k_n)}\right).$$

Now, the question is how quickly a possible change-point k_n^* can be detected, that is, what can be said about the behaviour of the stopping times $\tau_n^{(1)}, \tau_n^{(2)}$ or the detection delays $\tau_n^{(1)} - k_n^*$, $\tau_n^{(2)} - k_n^*$ under the alter-

native H_1 ? In the next section it will turn out that the limiting distributional behaviour of the stopping times, suitably normalized, is an asymptotically normal one.

3. Asymptotics of stopping times under the alternative

Similar to Aue et al. [2008], we consider an "early change" scenario here, that is, we assume that the change-point k_n^* does not occur too late compared to the length of the training period in the following technical sense:

$$k_n^* = O\big(h_n \log^\gamma (n / h_n)\big) \text{ for some } \gamma > 0. \tag{1}$$

Theorem 3 (Gut and Steinebach [2008]). Assume Condition (1). If $\{h_n\}$ is as in Theorem 1, then, under the alternative H_1,

$$\frac{\tau_n^{(1)} - k_n^*}{\frac{\eta}{|\theta^* - \theta|}\sqrt{h_n}} - c_n^{(1)} \to N(0,1) \text{ in distribution (as } n \to \infty).$$

For the second stopping time we similarly assume

$$k_n^* = O\big(k_n \log^\gamma (n / k_n)\big) \text{ for some } \gamma > 0. \tag{2}$$

and have

Theorem 4 (Gut and Steinebach [2008]). Assume Condition (2). If $\{k_n\}$ is as in Theorem 2, then, under the alternative H_1,

$$\frac{\tau_n^{(2)} - k_n^*}{\frac{\eta}{|\theta^* - \theta|}\sqrt{k_n^*}} - c_n^{(2)} \to N(0,1) \text{ in distribution (as } n \to \infty).$$

Remark 2. Similar results as in Theorems 3 and 4 can also be obtained, if the in-control parameters θ, η and the drift coefficient θ^* after the change-point k_n^* are replaced by suitable estimates. For details we refer to Gut and Steinebach [2008], Theorems 6.2 and 7.2. As an immediate consequence one can construct an asymptotic confidence interval for the unknown change-point k_n^* from the stopping times $\tau_n^{(1)}$ and $\tau_n^{(2)}$, respectively (see Gut and Steinebach [2008], Corollaries 7.1-7.2).

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 7 (13)

References

Aue A., Horváth L., Kokoszka, P., Steinebach J., *Monitoring shifts in mean: Asymptotic normality of stopping times*. TEST, 17 (2008), pp. 515-530.

Gut A., Steinebach J., *Truncated sequential change-point detection based on renewal counting processes*. Scandinavian Journal of Statistics, 29 (2002), pp. 693-719.

Gut A., Steinebach J., *Truncated sequential change-point detection based on renewal counting processes II*. Journal of Statistical Planning and Inference, 2008, 16 pp. (online available: DOI 10.1016/j.jspi.2008.08.021).

ANALYSIS OF DEPENDENT RISKS

Stanisław Heilpern (Wrocław University of Economics)

1. Process with dependent claims

We will study the following risk process:

$$U(t) = u + ct - \sum_{i=1}^{N(t)} X_i,$$

where $U(t)$ is a surplus of the insurer at time t, u is an initial surplus, c is a premium rate and $N(t)$ is a counting Poisson process with intensity λ. Let us assume, that the claims $X_i, i = 1, 2, \ldots$, are independent of $N(t)$, but the claims may be dependent. We will investigate the probability of ruin of such process: $\psi(u) = P(T < \infty | U(0) = u)$, where $T = \inf\{t: U(t) < 0\}$ is time of ruin. The symbol $\psi_I(u)$ denotes the ruin probability for independent claims X_i and $\psi_{sd}(u)$ for the strict dependent case, when $X_i = X$ is the same random variable for every i. We obtain the following relations:

$$\psi_{sd}(0) \le \psi_I(0) \qquad \psi_{sd}(\infty) \ge \psi_I(\infty).$$

If $F\left(\frac{c}{\lambda}\right) < 1$ then we have strong inequalities.

Now, we assume, that the dependent structure of X_iXi is describe by the Archimedean copula C

$$\overline{F}(x_1, \ldots, x_n) = C(\overline{F}_1(x_1), \ldots, \overline{F}_n(x_n)) = \\ = g^{-1}(g(\overline{F}_1(x_1)) + \cdots + g(\overline{F}_n(x_n))),$$

where $g: [0,1] \to \mathrm{R}_+$ is the decreasing, completely monotonic, $g(0) = \infty$, $g(1) = 0$ function called generator and $\overline{F}(x) = 1 - F(x)$ is survival function. Then there exists the random variable Θ with distribution function

F_Θ such, that $M_\Theta(s) = g^{-1}(-s)$, where $M_\Theta(s)$ is a moment generating function of Θ, [Frees, Valdez 1998]. The claims X_i are conditional independent for $\theta \in \Theta$ in this case. So, we obtain the classical risk process $U_\theta(t) = u + ct - \sum_{i=1}^{N(t)} X_{i|\theta}$ with independent claims $X_{i|\theta}$ for fixed $\theta \in \Theta$. The unconditional probability of ruin is the mixture $\psi(u) = \int_0^\infty \psi_\theta(u) dF_\Theta(\theta)$ in this case, where $\psi_\theta(u)$ is the conditional ruin probability for fixed θ.

Let $m(\theta) = E(X_{i|\theta})$ be the expected value function. Then, $\psi_\theta(u) = 1$ for $\theta \le \theta_0$, where the border value θ_0 is a solution of equation $m(\theta_0) = \frac{c}{\lambda}$ and we obtain, that the unconditional probability of ruin is equal

$$\psi(u) = \int_{\theta_0}^\infty \psi_\theta(u) dF_\Theta(\theta) + F_\Theta(\theta_0).$$

Moreover $\psi(0) = \frac{\lambda}{c} \int_{\theta_0}^\infty m(\theta) dF_\Theta(\theta) + F_\Theta(\theta_0)$ and $\psi(\infty) = F_\Theta(\theta_0)$.

Theorem 1. $\psi(0) \le \psi_I(0)$ and $\psi(\infty) \ge \psi_I(\infty)$. If $F_\Theta(\theta_0) > 0$ then we obtain strong inequalities.

Example 1. Let us assume, that the dependent structure of X_i is describe by the Clayton copula $C_\alpha(u_1, \ldots, u_n) = (u_1^{-\alpha} + \cdots + u_n^{-\alpha} - n + 1)^{-\frac{1}{\alpha}}$, where $\alpha > 0, c = 24, \lambda = 4$ and the claims X_i have Pareto distribution $\overline{F}(x) = \left(\frac{3}{x+3}\right)^3$. The parameter α reflects the degree of dependence and Kendall-τ correlation coefficient is a simple function of it: $\tau = \frac{\alpha}{\alpha + 2}$. The induced random variable Θ has the gamma distribution $\mathrm{Ga}\left(\frac{1}{\alpha}, \alpha\right)$ and conditional survival function of claim is equal $\overline{F}_\alpha(x|\theta) = \exp\left(\frac{\theta}{\alpha}\left(1 - \left(\frac{x+3}{3}\right)^{2\alpha}\right)\right)$. The border value θ_α is solution of equation $e^{\frac{\theta}{\alpha}} \Gamma\left(\frac{0.5}{\alpha}, \frac{\theta}{\alpha}\right)\left(\frac{\alpha}{\theta}\right)^{\frac{1}{2\alpha}} = 4\alpha$, where $\Gamma(a,b) = \int_b^\infty x^{a-1} e^{-x} dx$. The values of ruin probabilities for different á and u are presented in the table 1.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

Table 1. The values of ruin probabilities for different α and u

α	u							
	0	4	20	60	100	200	400	600
0	0.5000	0.3071	0.1376	0.0546	0.0310	0.0160	0.0070	0.0050
2/3	0.3803	0.2581	0.2033	0.1627	0.1435	0.1204	0.1011	0.0955
2	0.3452	0.2285	0.1980	0.1739	0.1583	0.1329	0.1122	0.1063
∞	0.3333	0.1778	0.1301	0.1181	0.1160	0.1120	0.1112	0.1111

Source: own calculations.

We see, that for smaller values of initial capital u, the greater degree of dependence implies the smaller probability of ruin. For the greater values of u we obtain the reverse relation, but for middle values, e.g. $u = 100$, the greatest probability of ruin is obtained by the middle degree of dependence.

2. Multidimensional process

Now, we will study the following multidimensional process

$$\begin{pmatrix} U_1(t) \\ \vdots \\ U_n(t) \end{pmatrix} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} + \begin{pmatrix} c_1 t \\ \vdots \\ c_n t \end{pmatrix} + \begin{pmatrix} \sum_{k=1}^{N(t)} X_{1k} \\ \vdots \\ \sum_{k=1}^{N(t)} X_{nk} \end{pmatrix}.$$

We assume, that the claims $X_{i1}, X_{i2}, \ldots$ are independent with cumulative distribution function (cdf.) $F_i(x)$ for every $i = 1, \ldots, n$ and the random vectors of claims $(X_{1k}, \ldots, X_{nk})$ have the same distribution for any k with joint cdf. $F(x_1, \ldots, x_n)$. The claims $X_{1k}, \ldots, X_{nk}$ may be dependent and $N(t)$ is a Poisson process with intensity λ.

This multidimensional process has the following interpretation. We have n different types of claims and every claim event (eg. accident in communication insurance) can induce various types of claims (eg. vehicle damage, personal injury).

We will investigate the sum of such processes:

$$U(t) = U_1(t) + \cdots + U_n(t) = u + ct - \sum_{k=1}^{N(t)} Z_k, \qquad (1)$$

where $u = \sum_{i=1}^{n} u_i$, $c = \sum_{i=1}^{n} c_i$ and $Z_k = \sum_{i=1}^{n} X_{ik}$. So, we obtain the classical risk process with independent aggregated claims Z_k, which have the same distribution.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

Then, we will study the impact of dependence of claims in each claim event on the probability of ruin. Let, $\mathbf{X}_k = (X_{1k}, \ldots, X_{nk})$ and $\mathbf{Y}_k = (Y_{1k}, \ldots, Y_{nk})$ be two random vectors with the same marginal distributions but different joint cdf. $F_{\mathbf{X}}, F_{\mathbf{Y}}$ and $\psi_{\mathbf{X}}(u), \psi_{\mathbf{Y}}(u)$ denote the probability of ruin for suitable risk process (1). We will use the supermodular order between such random vectors to investigate this impact. The random vector $\mathbf{X}$ is smaller than the random vector $\mathbf{Y}$ in the supermodular ordering, written $\mathbf{X} \leq_{\text{sm}} \mathbf{Y}$, or $F_{\mathbf{X}} \leq_{\text{sm}} F_{\mathbf{Y}}$, if $E(f(X)) \leq E(f(Y))$ for all supermodular functions $f: R^n \to R$, i.e. $f(\mathbf{x}) + f(\mathbf{y}) \leq f(\mathbf{x} \vee \mathbf{y}) + f(\mathbf{x} \wedge \mathbf{y})$ for any $\mathbf{x}, \mathbf{y} \in R^n$ [Shaked, Shanthikumar 1997].

Theorem 2 [Cai, Li 2005]. If $\mathbf{X} \leq_{\text{sm}} \mathbf{Y}$ then $\psi_{\mathbf{X}}(u) \leq \psi_{\mathbf{Y}}(u)$.

Let, the dependence structure of random vector $\mathbf{X}$ is described by copula $C_{\mathbf{X}}$. The copula is joint cdf. of uniform random variables, so $\mathbf{X} \leq_{\text{sm}} \mathbf{Y} \Leftrightarrow C_{\mathbf{X}} \leq_{\text{sm}} C_{\mathbf{Y}}$, because the supermodular order is closed under monotonic functions.

Now, we will investigate three cases.

a) The dependent structure or random vector $\mathbf{X}$ is described by **Archimedean copula** with generator $g_{\mathbf{X}}$. If $g_{\mathbf{X}} \circ g_Y^{-1} \in L_\infty^*$ then $C_{\mathbf{X}} \leq_{\text{sm}} C_{\mathbf{Y}}$, where $L_\infty^* \{w: [0,\infty) \to [0,\infty) \mid w(0) = 0, \quad w(\infty) = \infty, \ (-1)^{i-1} w^{(i)}(t) \geq 0, \quad i = 1, 2, \ldots\}$ [Wei, Hu 2002]. The families of Archimedean copulas C_α, e.g. Clayton, Frank or Gumbel, characterized by parameter á reflected the degree of dependence, are often used in practice. We often obtain relation, that $\alpha \leq \beta$ implies $C_\alpha \leq_{\text{sm}} C_\beta$ in this case. So, we have $\psi_\alpha(u) \leq \psi_\beta(u)$ from theorem 2.

b) The dependence is described by **elliptical copulas**, i.e. copulas induced by elliptically contoured distribution (Gauss, *t*-distribution, logistic).

Theorem 3 [Wei, Hu 2002; Müller 2001]. Let $\mathbf{X}$, $\mathbf{Y}$ be elliptically contoured distribution and $\sigma_{ij}^{\mathbf{X}} = Corr(X_i, X_j)$. If $\sigma_{ij}^{\mathbf{X}} \leq \sigma_{ij}^{Y}$ then $\mathbf{X} \leq_{sm} \mathbf{Y}$.

So, we obtain $\psi_{\mathbf{X}}(u) \leq \psi_{\mathbf{Y}}(u)$. We can generalize this fact on the case of elliptical copulas.

c) Extreme cases. Frechet space $\mathsf{R}_n(F_1, \ldots, F_n)$ or $\mathsf{R}_n(\mathbf{X})$, where $F_1, \ldots, F_n$ are the marginal cdf. of random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is a class of all joint distribution functions with the same marginal cdf. $F_1, \ldots, F_n$. We assume that $X_i \geq 0$. Every member F of this class is bounded by two functions M_n and W_n called the Frechet lower and upper bound:

$$M_n(x_1, \ldots, x_n) \leq F(x_1, \ldots, x_n) \leq W_n(x_1, \ldots, x_n).$$

The upper bound W_n is cdf. [Dhaene, Denuit 1999]. Then $W_n(x_1, \ldots, x_n) = \min\{F_1(x_1), \ldots, F_n(x_n)\}$ and the marginal random variables $X_1, \ldots, X_n$ are strict dependent, called comonotonic, in this case.

The lower Frechet bound is equal $M_n(x_1, \ldots, x_n) = \max\left\{\sum_{i=1}^{n} F_i(x_i) - n + 1,0\right\}$. For $n = 2$ the lower bound M_2 is a joint cdf. of random vector $\mathbf{X}^{\mathrm{L}}$ [Dhaene, Denuit 1999]. The marginal random variables X_1, X_2 called countermonotonic are strict dependent too, but they are reverse dependent. We obtain the following relations: $\mathbf{X}^{\mathrm{L}} \leq_{sm} \mathbf{X} \leq_{sm} \mathbf{X}^{\mathrm{U}}$, where $\mathbf{X}^{\mathrm{L}}, \mathbf{X}^{\mathrm{U}} \in \mathsf{R}_n(\mathbf{X})$ in this case, so we have

$$\psi_{\mathrm{L}}(u) \leq \psi_{\mathrm{X}}(u) \leq \psi_{\mathrm{U}}(u).$$

For $n > 2$, the situation is more complicated. The lower Frechet bound may not be cdf. in this case [Dhaene, Denuit 1999]. The random vector $\mathbf{X}^{\mathrm{m}} = (X_1, \ldots, X_n)$ is mutually exclusive if $P(X_i > 0, X_j > 0) = 0$ for $i \neq j$.

Theorem 4 [Dhaene, Denuit 1999; Bäuerle, Müller 1998]. If $\mathbf{X}^{\mathrm{c}}, \mathbf{X}^{\mathrm{m}} \in \mathsf{R}_n(\mathbf{X})$, where $\mathbf{X}^{\mathrm{c}}$ is upper Frechet bound and $\mathbf{X}^{\mathrm{m}}$ is mutually exclusive, then $\mathbf{X}^{\mathrm{m}} \leq_{sm} \mathbf{X} \leq_{sm} \mathbf{X}^{\mathrm{U}}$.

So we obtain from theorem 4, that

$$\psi_{\mathrm{m}}(u) \leq \psi_{\mathrm{X}}(u) \leq \psi_{\mathrm{U}}(u).$$

The mutually exclusive random vector $\mathbf{X}^{\mathrm{m}}$ may not exist, too. It exists iff the Frechet family $\mathsf{R}_n(F_1, \ldots, F_n)$ satisfies condition $n - 1 \leq F_1(0) + \ldots + F_n(0)$. If Frechet family satisfies such conditions, then the joint cdf. F is mutually exclusive iff it is lower Frechet bound [Dhaene, Denuit 1999].

Example 2. Let $n = 2, c_1 = c_2 = 4, \lambda = 1$ and the claims X_{1i}, X_{2i} have the exponential distributions with means $m_1 = m_2 = 1$. We investigate four cases:

a) Random vector (X_{1i}, X_{2i}) is comonotonic. Then the aggregate claims Z_i have exponential distribution with mean $m = 2$ and $\psi_c(u) = 0.25e^{-3u/8}$.

b) The dependence structure (X_{1i}, X_{2i}) is describe by Clayton copula with $\alpha = 2$, i.e. $\tau = 0.5$. We must derive probability of ruin using the numerical methods in this case.

c) Random variables (X_{1i}, X_{2i}) are independent. Then the aggregate claims Z_i have gamma distribution $G(2,1)$ and $\psi_I(u) = 0.321e^{-0.579u} - 0.071e^{-1,297u}$.

d) Random vector (X_{1i}, X_{2i}) is countermonotonic Then the aggregate claims Z_i have the following distribution function

$F_{Z_i}(z) = \begin{cases} 0 & z \le -\ln(0.25) \\ \sqrt{1-4e^{-z}} & z > -\ln(0.25) \end{cases}$. We must derive probability of ruin using the numerical methods in this case.

The values of ruin probabilities for the different cases and different values of initial capital *u* are presented in the table 2.

Table 2. The values of ruin probabilities for the different cases and *u*

u	Countermonotonic	Independent	Clayton $\tau = 0.5$	Comonotonic
0	0.25	0.25	0.25	0.25
4	0.0154	0.0313	0.0426	0.0558
8	0.0012	0.0031	0.0071	0.0124
12	0.0003	0.0003	0.0027	0.0028

Source: own calculations.

We obtain the more regular situation than in the example 1. If the dependence increases, then the probability of ruin increases for every value of initial capital *u*.

References

Bäuerle N., Müller A., *Modeling and comparing dependences in multivariate risk portfolios*. ASTIN Bulletin, 28 (1998), pp. 59-76.

Cai J., Li H., *Multivariate risk model of phase type*. Insurance: Mathematics and Economics, 36 (2005), pp. 137-152.

Dhaene J., Denuit M., *The safest dependence structure among risks*. Insurance: Mathematics and Economics, 25 (1999), pp. 11-21.

Frees E.W., Valdez E., *Understanding relationships using copulas*. North Amer. Actuarial J., 2 (1998), pp. 1-25.

Müller A., *Stochastic ordering of multivariate normal distributions*. Ann. Inst. Statist. Math., 53 (2001), pp. 567-575.

Shaked M., Shanthikumar J.G., *Supermodular stochastic orders and positive dependence of random vectors*. J. Multivariate Anal., 61 (1997), pp. 86-101.

Wei G., Hu T., *Supermodular dependence ordering on a class of multivariate copulas*. Insurance: Mathematics and Economics, 57 (2002), pp. 375-385.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

SOME PROPERTIES CONCERNING MISSING DATA ANALYSIS

Karlheinz Fleischer (Philipp University of Marburg)

1. Introduction

In surveys people don't answer all questions, usually. There is a variety of reasons for missing values. Most statistical methods are designed for data sets without missing values. Especially older software ignores missing values, but the fact that a value is missing might be informative. Hence, it is very important to consider techniques for missing data and to explore properties of such methods.

2. Notations

Let $Y = (Y_1, \ldots, Y_m)$ be a vector of variables of interest with values $(y_1, \ldots, y_m)$ and R a (also vector valued) response variable, i.e. $R_i = 1$, if the value of Y_i is observed, $R_i = 0$ otherwise ($i = 1, \ldots, m$).

Sometimes we will demonstrate some properties for two univariate variables, which will be denoted by X and Y and the response variables by R_X and R_Y (resp.).

Usually Y is separated in two subvectors Y_{obs} and Y_{mis}, where Y_{obs} contains the observed variables and Y_{mis} the unobserved variables for the actual sample.

P_W denotes the probability distribution of a (univariate or vector valued) random variable W.

3. Missing data mechanisms

Rubin distinguishes three types of missing data mechanisms [Little, Rubin 2002; Rubin 1976]:

Missing completely at random (MCAR):

$P_{R|Y_{obs},Y_{mis}}(1|y_{obs}, y_{mis}) = P_R(1)$ for all possible values y_{obs}, y_{mis} of Y_{obs}, Y_{mis}.

Missing at random (MAR):

$P_{R|Y_{obs},Y_{mis}}(1|y_{obs}, y_{mis}) = P_{R|Y_{obs}}(1|y_{obs})$ for all possible values y_{mis} of Y_{mis}.

Missing not at random/Not missing at random (MNAR or NMAR)
if $P_{R|Y_{obs},Y_{mis}}(1|y_{obs}, y_{mis})$ depends on y_{mis}.

Observe: The variables Y_{obs} and Y_{mis} change from unit to unit (or sample to sample)!

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

4. Some Remarks

According to Rubin, these conditions must only be satisfied for the actually observed missing data pattern R (see e.g., [Schafer, Graham 2002, p. 151]).

They argue that if the condition is satisfied for the actual sample one can make inferences in the same manner as for a sample without missing values. Otherwise, our interest is not to make inferences for one specific sample but we want to know how accurate specific methods (imputation procedures) will work. Hence we have to know what to do for every possible sample, not only for the actual one. We will assume that both (sets of) variables X and Y may be missing sometimes. In order to derive properties of certain missing data procedures we need general assumptions on the response probability for X and Y.

Inferences are usually based on the distribution $P_{Y_1,\ldots,Y_k}$. But according to missing values we do not observe this distribution. Hence, we will study those distributions which are observable. For simplicity we will assume 2 discrete random variables X, Y with response variables R_X, R_Y respectively.

5. Observable Distributions

The term observable distribution shall denote distributions which can be evaluated by taking samples subsequently. Hence, observable are the following distributions and probabilities:

- $P_{R_X,R_Y}(1,1)$ and its marginal and conditional probabilities (e.g. $P_{R_X}(1)$, $P_{R_X|R_Y}(1|1), P_{R_Y|R_X}(0|1)$, $P_{R_Y|R_X}(1|0), \ldots$),
- $P_{X,Y|R_X,R_Y}(x,y|1,1)$ and its marginal and conditional distributions, (e.g. $P_{X|R_X,R_Y}(x|1), P_{X|Y,R_X,R_Y}(x|y,1,1), \ldots$),
- $P_{X|R_X}(x|1), P_{Y|R_Y}(y|1)$,
- $P_{X,Y,R_X,R_Y}(x,y,1,1) = P_{R_X,R_Y}(1,1)\cdot P_{X,Y|R_X,R_Y}(x,y|1,1)$.

6. Some Properties

If $P_{R_X|X,Y}(1|x,y) = P_{R_X}(1)$ holds for all (x,y) within the support of (X,Y) we have

$$P_{R_X|X}(1|x) = P_{R_X|Y}(1|y) = P_{R_X|X,Y}(1|x,y) = P_{R_X}(1). \qquad (1)$$

Sketch of the proof: Since

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 7 (13)

$$P_{R_X|X}(1|x)=\frac{P_{R_X|X}(1,x)}{P_X(x)}=\sum_y\frac{P_{R_X|X,Y}(1,x,y)}{P_X(x)}=$$

$$=\frac{1}{P_X(x)}\sum_y P_{R_X|X,Y}(1|x,y)P_{X,Y}(x,y)=$$

$$=\frac{1}{P_X(x)}\sum_y P_{R_X}(1)P_{X,Y}(x,y)=\frac{P_{R_X}(1)}{P_X(x)}\sum_y P_{X,Y}(x,y)=$$

$$=\frac{P_{R_X}(1)}{P_X(x)}P_X(x)=P_{R_X}(1)=P_{R_X|X,Y}(1|x,y).$$

$$P_{R_X|Y}(1|y)=\frac{P_{R_X|Y}(1|y)}{P_Y(y)}=\sum_x\frac{P_{R_X,X,Y}(1,x,y)}{P_Y(y)}=$$

$$=\frac{1}{P_Y(y)}\sum_x P_{R_X|X,Y}(1|x,y)P_{X,Y}(x,y)=$$

$$=\frac{1}{P_Y(y)}\sum_x P_{R_X}(1)P_{X,Y}(x,y)=\frac{P_{R_X}(1)}{P_Y(y)}P_Y(y)=P_{R_X}(1).$$

Furthermore, it holds according to Bayes formula

$$P_{X,Y|R_X,R_Y}(x,y|1,1)=P_{X,Y}(x,y)\frac{P_{R_X,R_Y|X,Y}(1,1|x,y)}{P_{R_X,R_Y}(1,1)}. \qquad (2)$$

From (2) we can conclude the (well-known) property

$$P_{X,Y|R_X,R_Y}(x,y|1,1)=P_{X,Y}(x,y) \quad \Leftrightarrow \quad P_{R_X,R_Y|X,Y}(1,1|x,y)=P_{R_X,R_Y}(1,1).$$

Additionally, it holds:

$$P_{R_X,R_Y|X,Y}(1,1|x,y)=P_{R_X|X,Y}(1|x,y)\cdot P_{R_Y|X,Y,R_X}(1|x,y,1) \qquad (3)$$

$$=P_{R_Y|X,Y}(1|x,y)\cdot P_{R_X|X,Y,R_Y}(1|x,y,1) \qquad (4)$$

⇒ Even if $P_{R_X|X,Y}=P_{R_X}$, we need a model for the response probability on Y depending on X,Y and on R_X and not only for the response probability conditional on X and Y alone.

Furthermore, it holds in general

$$P_{X|R_X}(x|1)=\sum_{i=0}^{1}P_{X,R_Y|R_X}(x,r|1)=\sum_{i=0}^{1}P_{R_Y|R_X}(r|1)\cdot P_{X|R_X,R_Y}(x|1,r)\neq$$

$$\neq\cdot P_{X|R_X,R_Y}(x|1,1).$$

That means that usually the observable distribution of X (i.e. under the condition $R_X = 1$) does not correspond to the observable distribution of X under the condition that X and Y are both observed.
But since

$$P_{X|R_X}(x|1) = P_{R_X|X}(1|x)\frac{P_X(x)}{P_{R_X}(1)}$$

we observe, that in case of $P_{R_X|X,Y} = P_{R_X}$ then $P_{X|R_X}(x|1) = P_X(x)$ (according to (1)) but $\cdot P_{X|R_X,R_Y}(x|1,1) \neq P_X(x)$ in general.
At the end let us have a look at what happens if missing values are imputed using hot deck imputation?
Let X^*, Y^* be the values of X, Y after the imputation (i.e. $X^* = X$, if X is not missing and similar for Y^*, Y). Then it holds for all (x, y)

$$P_{X^*,Y^*}(x,y) = P_{X,Y}(x,y)\cdot P_{R_X,R_Y|X,Y}(1,1|x,y)\cdot$$
$$\cdot\left(1+\frac{P_{X,R_X,R_Y}(x,1,0)}{P_{X,R_X,R_Y}(x,1,1)}+\frac{P_{Y,R_X,R_Y}(x,1,0)}{P_{X,R_X,R_Y}(x,1,1)}+\frac{P_{R_X,R_Y}(0,0)}{P_{R_X,R_Y}(1,1)}\right).$$

Assuming different missing data mechanisms leads to certain relations which may be presented in another talk.

References

Little R.J.A., Rubin D.B., *Statistical Analysis with Missing Data*. Wiley, New York 2002.

Rubin D.B., *Inference and missing data*. Biometrika, 63 (1976), pp. 581-592.

Schafer J.L., Graham J.W., *Missing data: our view of the state of the art*. Psychological Methods, 7 (2002), pp. 147-177.

THE DEMOGRAPHIC CHARACTERISTICS AND THE RELATIVE ECONOMIC STATUS OF FAMILIES IN POLAND. EMPIRICAL RESEARCH

Zofia Rusnak (Wrocław University of Economics)

The problems of evaluating and comparing the economic status of various groups of households (especially biological families consisting of child-

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

less marriages and marriages with different number of children) have been the subject of my research for several years.

The main aim of this work is to present the empirical research results related to determining relative income or expenses needs in given family types.

The explanation of the idea of relative income (or expenses) can be fond in the basic question: "What income should be at the disposal of a family with children in comparison to the income of a marriage without children so that both of these family types are equally wealthy?" .

The answer is possible when we define and estimate the equivalence scales, which are the multipliers scaling the incomes and expenses of families of different demographic profile making them liable to comparison.

There are two significantly different approaches used for the estimation of the equivalence scales: objective – embracing the so called normative and empirical methods of defining scales; and subjective – where the scales are based on the data collected from the households and related to the households' own perspective on various levels of income.

Among the normative scales, the most commonly used – especially in public statistics in EU countries (mainly in the analyses of income inequality and poverty range) – are the OECD scales with two parameters, calculated as follows:

$$m_{\alpha/\beta} = 1 + \alpha(n_{\alpha} - 1) + \beta \cdot n_{c},$$

where n_a and n_c stand for the number of adults and number of children in the household respectively, while α and β are arbitrarily set parameters.

In the original (standard) OECD scale type 70/50 $\alpha = 0.7$ and $\beta = 0.5$, which means that according to this scale the coefficient equals 1 for the first adult person, 0.7 for the next adult, and 0.5 for every child. In developed EU countries the so called modified OECD scale type 50/30 is becoming more and more often used. Its usage results from the decrease in the share of food expenses in the budget of households in these countries.

In accordance with recommendation of EUROSTAT in Poland since 2005 in public statistics the modified scale is used in analyses of income inequality and poverty sphere.

Various methods are used to estimate empirical equivalence scales – one of the oldest of which while at the same time most often employed is the Engel's method. In this method one needs to define the formula of the Engel's curve which describes the relationship between the share of expenses on food in all expenses and various social-economic and demographic characteristics of a household.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

In this work there were following basic research issues connected with defining Engel's curves:

- a choice of variables describing the demographic profile of a family,
- a way of using these variables in Engel's curve,
- a choice of Engel's curve which matches the empirical data the best.

The statistical data used for all calculations was:

- aggregate data related to income and expenses (per capita) from Polish CSO publications from 1993 to 2004,
- unit data from household budget research carried out by CSO in Poland in 2004.

In case of aggregate data, when the only information available is about the number of people in the family and the number of children, the following formulas for the Engel's curve have been applied:

$$w_t^k = \alpha + \beta \ln(x_t^k) + \eta \ln n_k + \gamma_1 r_1^k + \vartheta_1 \ln p_t + \vartheta_2 \ln p_{gt} + \\ + \tau_1 z_1 + \varepsilon, \tag{1}$$

$$w_t^k = \alpha + \beta \ln(x_t^k) + \eta \ln(1 + n_c^k) + \vartheta_1 \ln p_t + \vartheta_2 \ln p_{gt} + \\ + \sum_{i=1}^{4} \tau_i z_i + \varepsilon, \tag{2}$$

$$w_t^k = \alpha + \beta \ln\left(x_t^k\right) + \eta \cdot n_c^k + \vartheta_1 \ln p_t + \vartheta_2 \ln p_{gt} + \sum_{i=1}^{4} \tau_i z_i + \varepsilon, \tag{3}$$

where w_t^k stands for the share of food expenses in all expenses, x_t^k stands for all income (or expenses) in the household with k children to keep (that is the k type of household), r_1^k is the quotient of adults and the number of people in the k type of household, n_c^k is the number of children in k type of household. The variables z_1, z_2, z_3 and z_4 are dummy variables which equal 1 for household consisting of a single mother with children (type M+), a marriage with one child (A1), with two children (A2) and with three children (A3) respectively and which equal 0 in other types of households. The parameters used in these formulas have been estimated on the basis of data from the years 1993-2004 by means of the least squares method. While the collected data is in form of time series, the global consumer price index p_{gt} and food price index p_t have been taken into consideration in the abovementioned formulas. These indices were originally chain indices and – for the purpose of comparison – have been transformed into fixed-base indices, for which the year 1993 has been fixed as the basic period.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

The estimated parameters of all three formulas are shown in table 1.

Table 1. Values of the Engel's curves' estimated parameters

Engel's curve from formula (1)									
Parameters	α	β	η	ϑ_1	ϑ_2	τ_1	γ_1		
$R^2 = 0{,}997$	0.92	–0.183	0.346	0.37	–0.234	0.127	0.379		
$\lvert t \rvert$	3.68	5.99	13.45	10.04	3.83	4.55	7.64		
Engel's curve from formula (2)									
Parameters	α	β	η	ϑ_1	ϑ_2	τ_1	τ_2	τ_3	τ_4
$R^2 = 0{,}992$	1.57	–0.188	0.088	0.333	–0.19	–0.098	–0.045	–0.482	–0.321
$\lvert t \rvert$	9.70	7.2	30.10	11.56	3.83	14.22	10.26	11.17	10.79
Engel's curve from formula (3)									
Parameters	α	β	η	ϑ_1	ϑ_2	τ_1	τ_2	τ_3	τ_4
$R^2 = 0{,}992$	1.607	–0.194	0.033	0.333	–0.182	–0.068	–0.016	–0.017	–0.009
$\lvert t \rvert$	9.97	7.45	30.45	11.66	3.71	11.26	2.88	3.42	2.88

Source: own calculations; the *p*-value did not exceed 0.006.

The recurring negative values of estimated β parameter and positive values of estimated η parameter, standing by the variables relating to the size of the household show that the increase in all expenses results in a decrease in the share of food expenses (in *ceteris paribus* conditions), while the growth of a household without an increase in expenses implies an increase in this share. The attempts to include variables z_2, z_3, and z_4 in the set of explanatory variables in the first formula have been unsuccessful; the parameters for these variables were not significantly different from zero. The equivalence scales have been calculated by employing the estimated parameters in the following formulas:

$$S_k = \left(\frac{n_k}{2}\right)^{\left(-\frac{\eta}{\beta}\right)} \cdot e^{\frac{\gamma_1}{\beta}\left(1-n_1^k\right)} \cdot e^{-\frac{\tau_1}{\beta}z_1}, \tag{4}$$

$$S_k = \left(1+n_c^k\right)^{\left(-\frac{\eta}{\beta}\right)} \cdot \exp\left\{-\sum_{i=1}^{4}\frac{\tau_i z_i}{\beta}\right\}, \tag{5}$$

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

$$S_k = \exp\left\{-\frac{\eta}{\beta} n_c^k - \sum_{i=1}^{4} \frac{\tau_i}{\beta} z_i\right\}. \quad (6)$$

The results of the calculations are included in table 2.

Table 2. Equivalence scales calculated by use of Engel's method

Biological family type	Equivalence scales calculated by use of the formula		
	(4)	(5)	(6)
A0- marriage without children	1.000	1.000	1.000
A1- mariage with 1 child	1.042	1.062	1.064
A2- mariage with 2 children	1.326	1.313	1.312
A3- mariage with 3 children	1.749	1.747	1.734
A4+ mariage with 4 or more children	2.690	2.659	2.643
M+ mother/father with children	0.929	0.938	0.937

Source: own calculations.

The results for the scale calculated by use of formula (4) should be interpreted as follows: the expenses (income) of marriages with one, two, three and at least four children, enabling them to attain the standard of living comparable to that of a childless marriage, should be higher by respectively 4.2%, 32.6%, 74.9%, 169%, while lower by 7.1% for a single mother with children. The interpretation of the scale values obtained by use of formulas (5) and (6) is analogous.

The results obtained by use of formulas (4), (5), and (6) are very similar. It is reflected by fig. 1., in which the values of the relative income indicator – a quotient of real income and equivalent income – are shown.

Independently of the used Engel's curve and obtained equivalence scales, it is only in marriages with one child that both real income and expenses are higher than the equivalent ones, which means that the income and expenditure situation in this family type is relatively better than in family type A0. In other types of families the real income was too low for those families to attain the material standard of a childless marriage. It is the families with at least three children that are in relatively the worst situation. In these families the real income was lower than the equivalent income by 26-66%. Similar results obtained by use of these scales are the main reason why only the scale calculated by means of formula (4) has been used for further analysis.

ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY

Nr 7 (13)

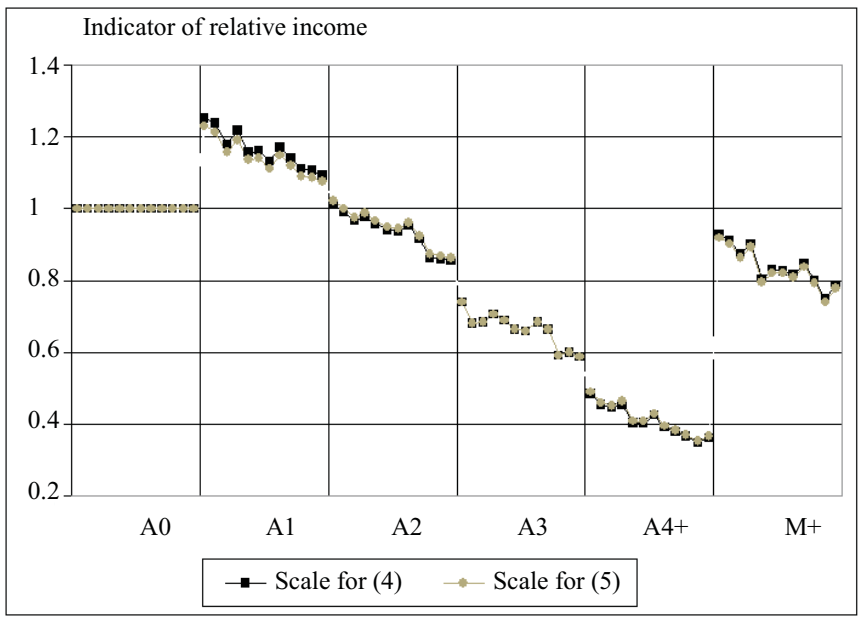

Fig. 1. Relative income indicator for different scales in years 1993-2004

Source: own calculations.

As far as the data from the household budget survey carried out in 2004 is concerned this formula was modified in two ways.

Firstly, as a result of dividing all the people ascribed to a given household into three groups – parents, children under 14, children over 14 – the parameters have been calculated for the Engel's curves defined in the following formulas:

$$w_{\dot{z}}^{k} = \alpha + \beta \ln\left(x^{k}\right) + \eta \ln n_{k} + \gamma_{1} r_{1}^{k} + \varepsilon \tag{7}$$

$$w_{\dot{z}}^{k} = \alpha + \beta \ln\left(x^{k}\right) + \eta \ln n_{k} + \gamma_{1} r_{1}^{k} + \gamma_{2} r_{2}^{k} + \varepsilon \tag{8}$$

r_{2}^{k} – the ratio of the number of children under 14 (marked by n_{c}) to the number of people in the household; D_{1} – dummy variable which equals 1 when there are children under 14 in the family, ε – random term; $r_{1}^{k} = 2 / n_{k}$ quotient in case of marriages and $1 / n_{k}$ in case of a single parent with children.

The general shape of equivalence scale calculated by use of the Engel's method on the basis of formulas (7), (8), and (9) is as follows:

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 7 (13)

$$S_k = \left(\frac{n_k}{2}\right)^{-\frac{\eta}{\beta}} \cdot \exp\left[\frac{\gamma_1}{\beta}\left(1 - r_1^{\,k}\right)\right] \cdot \exp\left(-\frac{\gamma_2}{\beta} r_2^{\,k}\right), \tag{9}$$

while the parameters γ_2, δ_1 equal 0 if in the given Engel's curve formula there are no variables related to these parameters.

The estimated parameters of given Engel's curves as well as respective equivalence scales are presented in tab. 3 and 4.

Table 3. Engel's curves' parameters estimations

Engel's curve for formula (7)					
Parametry	α	β	η	γ_1	
$R^2 = 0{,}512$	1.488	–0.188	0.173	0.113	
$\|t\|$	110.6	131.8	37.27	16.57	
Engel's curve for formula (8)					
Parametry	α	β	η	γ_1	γ_2
$R^2 = 0{,}516$	1.509	–0.189	0.182	0.096	–0.048
$\|t\|$	111.25	132.62	38.61	13.67	10.51

Source: own calculations.

At the significance level $\alpha = 0.01$ all estimated parameters in the above-mentioned formulas were statistically significant.

Table 4. Equivalence scales for given family types in 2004

Biological family type	Scales calculated on the basis of Engel's curves for formulas					
	(7)	(8)				
		$nc = 0$	$nc = 1$	$nc = 2$	$nc = 3$	$nc = 4$
A1	1.189	*1.246*	**1.146**			
A2	1.403	*1.509*	**1.417**	***1.331***		
A3	1.622	*1.777*	**1.690**	***1.607***	***1.528***	
A4	1.844	*2.046*	**1.962**	***1.881***	***1.804***	***1.731***
A4+	1.979	*2.210*	**2.128**	***2.048***	***1.972***	***1.899***
M+1	0.740	*0.776*	**0.685**			

Source: own calculations.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

In every calculated scale the higher number of children is matched by a higher equivalence scale value, which means that for such families the relative income needs which would allow the attainment of the economic status of a childless marriage are higher.

When only the *r*1 variable is used in the Engel's curve formula, the calculated scale is the most similar to the modified OECD scale 50/30. The increments in the scale show that the income of the families with a larger number of children should be higher than the income of a childless marriage by 19-22% (20% in the modified OECD scale).

Introduction of other variables, which define the demographic profile of a family in greater detail results in the flattening of the scale especially for families with one child (type A1 and M+1). Relative income needs resulting from the use of formula (8) amount to 14.6-20.3% of a marriage without children income, when all the children in the family are under 14. In the case when all the children are over 14, the highest values of the scale have been calculated. The relative income needs for every next child amount to over 25% for a family with one child and almost 28% for a family with four children (29.4% in the original OECD scale).

These results show that the equivalence scale's values depend a lot on the variables which describe the demographic structure of the family and on the character of statistical data used in calculations.

The largest differences can be found in family types A1 and A4+ especially when children are under 14.

References

Budżety gospodarstw domowych według wybranych typów rodzin w latach 1993-1996, GUS, Warszawa 1997.

Rusnak Z., *Statystyczna analiza dobrobytu ekonomicznego gospodarstw domowych*, AE, Wrocław 2007.

Warunki życia ludności w 1997 r., GUS, Departament Statystyki Społecznej, Warszawa 1998.

Warunki życia ludności w 1999 r., GUS, Departament Statystyki Społecznej, Warszawa 2000.

Warunki życia ludności w 2001 r., GUS, Departament Statystyki Społecznej, Warszawa 2002.

Warunki życia ludności w 2002 r., GUS, Departament Statystyki Społecznej, Warszawa 2003.

Warunki życia ludności w 2004 r., GUS, Departament Statystyki Społecznej, Warszawa 2006.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

HOW ECONOMIC MODELING MAY HELP WITH MNAR DATA

Christian Westphal (University of Marburg)

1. Motivation

Reviewing current literature a few things can be found out about missing data: Missing data is a present problem. There is very recent literature (namely [McKnight et al. 2007]) trying to spread the word to applied researchers. Combining this with the knowledge of increasing amounts of missing data in all sorts of surveys (see e.g. [de Leeuw, de Herr 2002; Vehovar et al. 2002, p. 233]) one can conclude, that missing data will be an increasing problem of the near future. A recent discussion with a good friend of mine, working in applied market research, has shown how flawed views on missing data can be. I was presented with the elimination of incomplete survey responses by not allowing them to be committed to the data base[1]. Obviously solutions like that are worse than accepting missingness. In this case it might have led to flawed responses or to unit nonresponse where item nonresponse could have been had.

Therefore we need easy to understand and easy to apply models for dealing with missing data. Dismissing the need for MNAR modeling [Schafer, Graham 2002, p. 20] contradicts my knowledge from the economics of information (see e.g. [Stiglitz 2000]).

2. Developing the Model

I started talking about the economics of information and I will use it now to develop a model for missingness in micro data. Most surveys ask for micro data, defined here as data revealing detailed information about survey/market participants. Data is information and information does hold a value [Allen 1990, esp. p. 271]. Now let us say data does hold a value for the surveyist but the revealing of the data may also hold a value for the participant. As an extreme example let us use your tax forms: The surveyed item shall be your yearly income and your deductibles. Now the lower your income and the higher your deductibles the higher your tax refund will be and vice versa (simplified of course). Given a low income and high deductibles you do have a severe incentive to respond (not responding might also be punishable, but even if it were not your incentive would still be there). No deductibles and a high income are not so favorable for the in-

[1] This was a web survey and the participant would not be able to finish without filling in all values.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

centive; this is where the punishment is needed to make people respond. Lying is not an option for the same reason.

When dealing with your common microeconomic survey the incentives for the participant are not as clear. Responding or not responding may lie very close together in terms of incentives. It will most likely be impossible for the participant to exactly know his value of responding to the survey. This can be seen as a direct extension to the decision model leading to response or nonresponse by [Beatty, Hermann 2002]. This model (with my extension) is illustrated in figure 1. A short explanation of the figure is as follows: On the first stage of the decision process the participant evaluates whether or not he can provide an answer. Data is either readily available (cost-free) or accessible (costs c go up) or generatable (costs go up even more) or not generatable (costs are infinite). On the next stage an adequacy judgment ("Do I find it adequate[2] to give the answer I have found in the earlier stage?") occurs. The last stage finally yields the latent variable binary choice model I am talking about: Based on some latent value of disclosing the surveyed item's value the participant makes his response decision[3]. Deviating from Beatty & Hermann's model I regard communicative intent as the outcome of the decision process and not as a decision stage as can be seen in figure 1. This outcome is either positive (the subject reveals its item value) or negative (the subject does not reveal its item value). Having talked of a value of responding earlier, I will

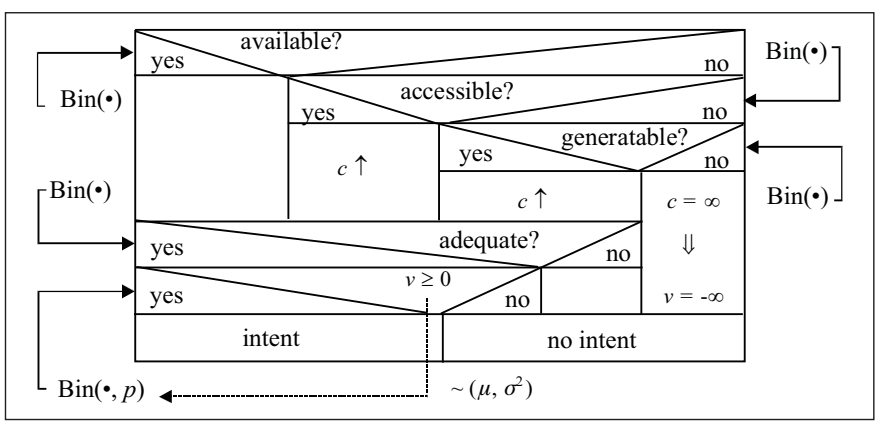

Fig. 1. Decision process leading to nonresponse

Source: own elaboration.

[2] E.g. adequate to some social rules.

[3] The binary choice random variables on each stage follow a Bernoulli distribution for one participant and abinomial distribution for $n > 1$ participants in the survey.

now use this value to explain the binary outcome of each individual's response decision. In case the participant expects his personal value of responding to be non-negative he will reply; else he will not reply.

So far this is a deterministic model. Your value of responding is non-negative, you reply; the value is negative, you do not. If this were true, for a group homogeneous in socioeconomic properties and utility function this meant, everyone would have to decide in the same way. This is not the case in reality. I can imagine several reasons:

- We simply ignore some random, unobserved differences between people.
- People are not capable of always correctly estimating their value of responding. Therefore among homogeneous people this estimated value differs randomly.

This randomness leads to a binary choice model known from econometrics [Greene 2008, chapter 23]. v from figure 1 is the latent variable. We shall regard it as a random utility level as used in random utility models. An outline of random utility models fitting my research idea very well can be found in [Boxall, Adamowicz 2002, p. 421-427].

With [Philipson 2001] there is an interesting article yielding evidence for my point of view and illustrating a way to learn about the missingness of data. Philipson used data of a survey among physicians asking for their annual income. To a randomly selected subsample an incentive of USD 50.00 was offered, should they complete the questionnaire. Basically this split the original sample ($n_{\text{total}} = 541$) into two samples, one being offered an incentive ($n_{\text{incentive}} = 243$)[4]. There were differences in the outcome of the survey. The response rate as well as the average income reported was significantly higher among the physicians being offered (and paid) an incentive. From my perspective developed earlier, this leads to conclusions about v among physicians:

- v can be shifted to the right[5].
- v is most likely negatively correlated with income.

I shall call the shift (or more generally: transformation) of $v, v^* = v^*(v, m)$ where m is the incentive offered to the participant.

3. First Results

As a first result I can show how unlikely it is for any missing data to be missing completely at random in the case of the model outlined above

[4] Actually there were incentives of two different sizes.

[5] The transformation actually may be more complex than a linear shift.

when we can distinguish between several heterogeneous groups of survey participants[6]. For MCAR to be fulfilled it has to hold

$$\int_{-\infty}^{0} f_{V_i}(v)dv = \int_{-\infty}^{0} f_{V_j}(v)dv \quad \forall(i,j). \tag{1}$$

Illustrated for groups i : $1 \rightarrow n$ with normally distributed V_i this condition can be simplified to:

$$\sigma_i = \mu_i \sigma_j / \mu_j \quad \forall(i,\ j). \tag{2}$$

Figure 2 illustrates how strong an assumption this would be. All cumulative distribution functions would have to meet in $F_{V_i}(v)$. As there is no reason to believe that distribution's parameters would behave as described in (1) and (2) we have to dismiss MCAR in general when dealing with micro data of socioeconomic distinguishable groups.

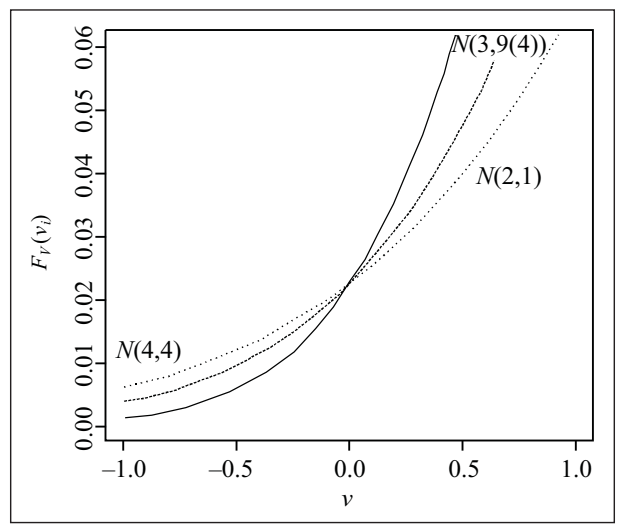

Fig. 2. Illustration for condition (2)

Source: own elaboration.

A more general result is the fundamental development of a model describing non-random missingness for a wide variety of data. From its design the model will help modeling the missingness in microeconomic data in a standardized way. Hereby the drawback in missing data research mentioned by [Rubin 1976, p. 589], that models for missingness "have not received much attention in the statistical literature" may be overcome instead of being swept aside as in e.g. [Schafer, Graham 2002, p. 154][7] and [Schafer 2003, p. 20][8].

[6] Distinguishing upon socioeconomic variables will be possible in most cases by sensible use of cluster analysis (see also [Boxall, Adamowicz 2002]).

[7] "... because in many psychological research settings the departures from MAR are probably not serious".

[8] "As a general principle, I believe that an analyst's time and effort are better spent building an intelligent model for the data rather than modelling the missingness, unless departures from MAR are suspected to be very serious".

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

4. Outlook

What still has to be done is to write down the model in a formal way in the best case reusing the notation of known random utility model literature and missing data literature.

Hopefully more theoretical results can be derived from the model. Of course extracting information from surveys is the main goal of the model formulation. I will have to see what information can be extracted. Recommendations for the design of surveys where MNAR data is expected may also be found.

Empirical studies have to show if my time used "modelling the missingness" [Schafer 2003, p. 20] will have been time well spent. The information gained by paying incentives may easily be used to verify/reject assumptions about the missingness.

Grouping (clustering) as in [Boxall, Adamowicz 2002] will hopefully tell us more about who is prone to nonresponse.

References

Allen B., *Information as an economic commodity*. New Developments in Economic Theory, 80(2), 1990, pp. 268-273.

Beatty P., Hermann D., *To Answer or Not to Answer: Decision Processes Related to Survey Item Nonresponse*, [in:] R.M. Groves, D.A. Dillman, J.L. Eltinge, R.J.A. Little (eds.), *Survey Nonresponse*, Wiley 2002.

Boxall P.C., Adamowicz W.L., *Understanding heterogeneous preferences in random utility models: a latent class approach*. Environmental and Resource Economics, 23(4), 2002, pp. 421-446.

de Leeuw E., de Heer W., *Trends in Household Survey Nonresponse: A Longitudinal and International Comparison*, [in:] R.M. Groves, D.A. Dillman, J.L. Eltinge, R.J.A. Little (eds.). *Survey Nonresponse*, Wiley 2002.

Greene W.H., *Econometric Analysis*. 6th edn., Pearson 2008.

Groves R.M., Dillman D.A., Eltinge J.L., Little R.J.A. (eds.), *Survey Nonresponse*, Wiley 2002.

McKnight P.E., McKnight K.M., Sidani S., Figueredo A. José, *Missing Data – A Gentle Introduction*, Guilford 2007.

Philipson T., *Data markets, missing data, and incentive pay*. Econometrica, 69(4), 2001, pp. 1099-1111.

Rubin D.B., *Inference and missing data*. Biometrika, 63(3), 1976, pp. 581-592.

Schafer J.L., *Multiple Imputation in Multivariate Problems. When the Imputation and Analysis Models Differ*. Statistica Neerlandica, 57(1), 2003, pp. 19-35.

Schafer J.L., Graham J.W., *Missing Data: Our View of the State of the Art*. Psychological Methods, 7(2), 2002, pp. 147-177.

Stiglitz J.E., *The contributions of the economics of information to the twentieth century economics*. The Quarterly Journal of Economics, 115(4), 2000, pp. 1441-1478.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 7 (13)

Vehovar V., Batagelj Z., Lozar M.K., Zaletel M., *Nonresponse in Web Surveys*, [in:] R.M. Groves, D.A. Dillman, J.L. Eltinge, R.J.A. Little (eds.). *Survey Nonresponse*, Wiley 2002.

TESTS FOR NORMALITY OF ERROR VARIABLES IN ARMA MODELS

Heiko Grönitz (University of Marburg)

A well-known Goodness-of-fit test is the Cramer-von-Mises test. This test is suitable for the test problem

$$H_0 : F = F_0 \text{ against } H_1 : F \neq F_0,$$

whereas F is assumed to be the continuous distribution function of independent and identically distributed random variables and F_0 denotes a given continuous distribution function. As test statistic the quantity $n \cdot \int_{-\infty}^{\infty} \left| F_n - F_0 \right| dF_0$ with the empirical distribution function F_n is used. The simple test problem above is usually of marginal relevance. It is more interesting to test

$$H_0 : F \in \mathsf{F} \quad \text{against } H_1 : F \notin \mathsf{F} \quad .$$

Here F denotes a whole parametric class of distribution functions. We are interested in the class $\mathsf{F} = \{F(\cdot,\sigma) = \Phi\left(\frac{\cdot}{\sigma}\right) : \sigma \in (0,\infty)\}$. That is we test the hypothesis "F is a centered normal distribution". For this test problem the well-known statistic

$$T = n \cdot \int_{-\infty}^{\infty} \left| F_n - F(\cdot,\hat{\sigma}_n) \right|^2 dF(\cdot,\hat{\sigma}_n),$$

whereas $\hat{\sigma}_n$ is the maximum likelihood estimator for the standard deviation of normal distributed variables can be used. Suppose that the hypothesis H_0 is true. Then $F = F(\cdot,\sigma_0)$ for one $\sigma_0 \in (0,\infty)$ holds. The estimator $\hat{\sigma}_n$ is consistent for σ_0 and $F(x,\cdot)$ is continuous for every $x \in \mathsf{R}$. Then for every $x \in \mathsf{R}$ the difference between $F_n(x)$ and $F(x,\hat{\sigma}_n)$ is "small" at least for "large" sample size n. Hence T attains smaller values under H_0 than under the alternative H_1. Note under the hypothesis T converges in distribution to a limit variable Z. To proof it a result by Durbin [Durbin 1973] is

decisive, whereas we confine ourselves to the special class F as above. He considered the sequence of empirical processes

$$\left(\sqrt{n}(F_n(x) - F(x,\hat{\sigma}_n))\right)_{x\in\mathbb{R}} .$$

Note that these stochastic processes have paths in the Skorohod space D, that is the paths are right continuous and have left limits. For such stochastic processes there is a theory of convergence in distribution, see Billingsley [Billingsley 1968]. Durbin has shown that the sequence of empirical processes converges in distribution. This convergence is to comprehend as convergence in the Skorohod space. Primarily by a continuous mapping theorem the convergence of T to Z follows and asymptotic critical values for the test could be calculated in the literature, compare with Hoermann [Hoermann 2007].

Let us now regard the independent and identically distributed error variables et in the famous ARMA time series models. Such error variables are always assumed as centered, that is they have expectation zero. Denote the distribution function of the e_t with F. We would like to test the normality of the errors. Formally we study the test problem

$$H_0: F \in \mathsf{F} = \left\{\Phi\left(\frac{\cdot}{\sigma}\right): \sigma \in (0,\infty)\right\} \text{ against } H_1: F \notin \mathsf{F} = \left\{\Phi\left(\frac{\cdot}{\sigma}\right): \sigma \in (0,\infty)\right\}.$$

Of course we can not use the previously mentioned test statistic. The reason is that we would have to compute the empirical distribution function and the estimator $\hat{\sigma}_n$ from data $e_1, \ldots, e_n$. However we are not able to observe the error variables. One can solve this problem by using suitable residuals $\overline{e}_1, \ldots, \overline{e}_n$, see Kreiss [Kreiss 1991]. These residuals are functions of the observed time series values and depend on the time series coefficients. If the coefficients are unknown they can be estimated with standard concepts, e.g. with the Yule-Walker method. Then one can calculate an empirical distribution function $\overline{F}_n$ Fn and a maximum likelihood estimator $\hat{\sigma}_n$ which both base on the residuals instead of $e_1, \ldots, e_n$. The idea is now to use

$$T_1 = n \cdot \int_{-\infty}^{\infty} (\overline{F}_n(x) - F(x,\hat{\sigma}_n))^2 \, dF(x,\hat{\sigma}_n)$$

as test statistic. To realize this idea one has to deal with the distribution of T_1 under the hypothesis. We have fully proofed that T_1 also converges in distribution to Z – the limit variable of the statistic T. To obtain this own result we have considered the sequence of stochastic processes

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

$$\alpha_n := \left(\sqrt{n}\left(\overline{F}_n(x) - F(x,\hat{\sigma}_n)\right)\right)_{x\in\mathsf{R}} .$$

We have shown that α_n converges in distribution to a certain Gaussian Process in the Skorohod space. Mainly by a continuous mapping theorem we can conclude that T_1 converges in distribution to the limit random variable Z. Quantiles of the distribution of Z can be calculated numerically and are tabled in [Hoermann 2007]. We can use these quantiles q_α as critical values for the test statistic T_1. This leads to the decision rule for the level $\alpha \in (0,1)$:

$$\text{Reject } H_0 \text{ iff } T_1 \text{ exceeds the quantile } q_{1-\alpha}.$$

Remember that the error variables e_t are centered. However we have to emphasize this additional information is not used in the statistic T_1. We expect a greater power of the test by using a statistic containing the further information.

As starting point for the construction of such an statistic we use the empirical likelihood concept introduced by Owen [1988; 1990] and continued by Qin/Lawless [Qin, Lawless 1994] as well as Zhang [Zhang 1997]. Suppose that we have observed realizations of independent and identically distributed centered variables $X_1,\ldots,X_n$ (with finite variance). The empirical likelihood method delivers an estimator $\widetilde{F}_n$ for the distribution function of X_i. Note that every random variable with distribution function $\widetilde{F}_n$ has expectation zero – a property that does not hold in general for the Empirical distribution function.

In this situation of observable variables X_i Genz [Genz 2004] has proofed a functional limit theorem for the processes

$$\left(\sqrt{n}\left(\widetilde{F}_n(x) - F(x,\hat{\sigma}_n)\right)\right)_{x\in\mathsf{R}} .$$

This means that he has shown the processes' convergence in distribution in the Skorohod space. One can conclude by a continuous mapping theorem that

$$S = \int_{-\infty}^{\infty} n \cdot \left(\widetilde{F}_n(x) - F(x,\hat{\sigma}_n)\right)^2 dF(x,\hat{\sigma}_n)$$

has a limit in distribution, denoted with W. The distribution of W is treated in [Hoermann 2007] again.

For our unobservable error variables we must use a version of $\widetilde{F}_n$ basing on the times series residuals. Denote this version with $\widetilde{\widetilde{F}}_n$. Have a look at T_1 again and replace $\overline{F}_n$ by $\widetilde{\widetilde{F}}_n$. Then define

$$T = \int_{-\infty}^{\infty} n \cdot (\widetilde{\widetilde{F}}_n(x) - F(x, \hat{\sigma}_n))^2 \, dF(x, \hat{\sigma}_n).$$

Another own result is that T_2 converges in distribution to W, too. Although one must use residuals which are not independent the distribution limit stays unchanged. To obtain the result we proofed a functional limit theorem for the residual processes

$$(\sqrt{n}(\widetilde{\widetilde{F}}_n(x) - F(x, \hat{\sigma}_n))_{x \in \mathbb{R}} \ .$$

Again by a continuous mapping theorem and some calculations the desired convergence of T_2 follows. Let $q_{1-\alpha}$ be the $(1-\alpha)$-quantile of the distribution of W. These quantities can be used as critical values and are charted in [Hoermann 2007] again. Then we have the decision rule for the level $\alpha \in (0,1)$:

$$\text{Reject } H_0 \text{ iff } T_2 \text{ exceeds the quantile } q_{1-\alpha}.$$

Hence we are able to test the normality of the errors with T_1 and T_2. The calculation of these statistics requires determining integrals with respect to normal distributions. However one can show with a transformation theorem from measure theory and some manipulations that T_1 and T_2 both simplify to finite sum which can be computed easily.

Finally we compare the power of T_1 and T_2 within a small simulation study. First we consider Laplace distributed errors. We choose different levels α and different lengths n of the time series. For every α/n-combination we make 1000 tests with T_1 and afterwards 1000 tests with T_2. As measure for the power the frequency of rejecting the wrong hypothesis is used. For each test we generate an ARMA(1,1) series, estimate the time series coefficients with the Yule-Walker technique and compare the statistic with the critical value.

In an analog way we examine the power of T_1 and T_2 in the situation of t-distributed errors.

The results are impressive. For Laplace distributed errors the power of T_2 is up to 40% greater than the power of T_1. For *t*-distributed error variables the power of T_2 is even up to 45% greater than the power of T_1. Summarizing we can obtain a manifestly greater power by applying the statistic T_2 which makes use of the errors' expectation zero.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

References

Billingsley P., *Convergence of Probability Measures*, Wiley, New York 1968.

Durbin J., *Weak convergence of the sample distribution function when parameters are estimated*. Ann. Statist., 1 (1973), pp. 279-290.

Genz M., *Anwendungen des Empirischen Likelihood-Sch¨atzers der Fehlerverteilung in AR(1)-Prozessen*, *Dissertation, Fachbereich Mathematik und Informatik, Physik, Geographie, Justus-Liebig-Universit¨at Giessen*, 2004.

Hoermann E., *The Cramer-von-Mises Test for Centered Distributions, Diploma thesis, Mathematisches Institut, Justus-Liebig-Universit¨at Giessen*, 2007.

Kreiss J.P., *Estimation of the distribution function of noise in stationary processes*, Metrika, 38 (1991), pp. 285-297.

Owen A., *Empirical likelihood ratio confidence intervals for a single functional*. Biometrika, 75 (1988), pp. 237-249.

Owen A., *Empirical likelihood confidence regions*, Ann. Statist., 18 (1990), pp. 90-120.

Qin J., Lawless J., *Empirical likelihood and general estimating equations*. Ann. Statist., 22 (199), pp. 300-325.

Zhang B., *Estimating a distribution function in the presence of auxiliary information*. Metrika, 46 (1997), pp. 221-244.

ESTIMATION OF THE RUIN PROBABILITY

Anna Nikodem (Wrocław University of Economics)

The ruin probability is an important parameter in the actuarial risk theory. In the paper there will be shown the estimators of this parameter in the classical risk theory in various scenarios with regard to what is assumed known and what is to be estimated. Let consider the risk process

$$U(t) = u + ct - \sum_{i=1}^{N(t)} X_i,$$

where u is the initial capital, c is the premium rate. The X represent the sizes of claim, which are assumed to be independent identically distributed with mean μ. The process $N(t)$N(t) is the Poisson process with intensity λ.

When all parameter are known, we can determine the ruin probability, which is defined as

$$\psi(u) = P\big(U(t) < 0 \text{ for some } t > 0\big),$$

and the survival probability over an infinite time horizon as

$$\Phi(u) = 1 - \psi(u) = P(\inf_{t>0} U(t) > 0).$$

To calculate the ruin probability, we can use the asymptotic approximation. In the classical risk model the ruin probability satisfies

$$\psi(u) \sim \frac{c - \lambda\mu}{\lambda\mu'(R) - c} e^{-Ru} \text{ for } u \to \infty,$$

where R is the adjustment coefficient, which is a positive solution of

$$1 + \frac{cR}{\lambda} = m_X(R),$$

where $m_X(R)$ is a moment generating function of r.v. X.

In practice one or more parameters are unknown and need to be estimated from the available data. In the first case, let μ and λ are known, data $X_1, X_2, \ldots, X_n$ are available. The ruin probability is estimated by [Csörgő, Teugels 1990]

$$\hat{\psi}(u) \sim \frac{c - \lambda\mu}{\lambda \cdot \hat{m}_n(\hat{R}_n) - c} e^{-\hat{R}_n u}.$$

Estimator of R is calculated from

$$1 + \frac{cr}{\lambda} = \hat{m}_n(r),$$

where

$$\hat{m}_n(r) = \frac{1}{n} \sum_{i=1}^{n} \exp(rX_i).$$

If the insurer can't determine λ and μ, but has only data from observation over time period $[0, T]$, the ruin probability is estimated by [Grandell 1991]

$$\hat{\psi}_T(u) \sim \frac{c - \hat{\lambda}_T \hat{\mu}_T}{\hat{\lambda}_T \cdot \hat{m}'_T(\hat{R}_T) - c} e^{-\hat{R}_T u},$$

where $\hat{R}_T$ is a positive solution of $1 + \frac{cr}{\hat{\lambda}_T} = \hat{m}_T(r)$. The empirical moment generating function, the intensity of the claim process and the mean of X are estimated respectively by

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 7 (13)

$$\hat{m}_T(r)=\frac{1}{N(T)}\sum_{i=1}^{N(T)}e^{rX_i},\ \hat{\lambda}_T(r)=\frac{N(T)}{T},\ \hat{\mu}_T(r)=\frac{1}{N(T)}\sum_{i=1}^{N(T)}X_i.$$

If the insurer can determine the intensity λ and the data $X_1, X_2, \ldots, X_n$ are available, we can use the formula for discrete claim amounts. When we assume that all observation have the same probability $p_j = P(X = x_i) = \frac{1}{n}$, the ruin probability can be estimated by

$$\hat{\psi}(u)=1-\frac{\theta}{1+\theta}\sum_{k_1,\ldots,k_n}(-z)^{k_1+\ldots+k_n}e^z\prod_{j=1}^{n}\frac{p_j^{k_j}}{k_j!},$$

where z is expressed by $z=\frac{\lambda}{c}(u-k_1x_1-\ldots-k_nx_n)_+$, $k_1, \ldots, k_n = 0,1,2,\ldots$

This method of calculating the ruin probability is difficult when the sample is large. Due to this fact it is better to estimate the claim distribution by the diatom distribution. The locations of the two atoms are

$$z_1=\bar{x}-x,\ z_2=\bar{x}+y.$$

Comparing the moment of diatomic distribution with the sample moment we get this probability

$$p_1=P(Z=z_1)=\frac{s^2}{s^2+x^2},\ p_2=P(Z=z_2)=\frac{x^2}{s^2+x^2},$$

where

$$x=\frac{\sqrt{\kappa^6+4s^6}-\kappa^3}{2s^2},\ y=\frac{\sqrt{\kappa^6+4s^6}+\kappa^3}{2s^2}.$$

In the next method, the ruin probability is estimated by plugging the estimator of unknown parameter to the Pollaczeck-Khinchine formula. This formula is as follows

$$\Phi(u)=1-\psi(u)=\sum_{n=0}^{\infty}\left(1-\frac{\lambda\mu}{c}\right)\left(\frac{\lambda\mu}{c}\right)^n\left(F_X^s\right)^{*n}(u),$$

where $F_X^s(u)=\frac{1}{\mu}\int_0^u(1-F(y))dy$ is the equilibrium distribution.

It is assumed that $\lambda\mu$ is known, F is unknown, and the data $X_1, X_2, \ldots, X_n$ are available. Then the equilibrium distribution is estimated by

$$\hat{F}_X^s(u) = \frac{1}{\bar{x}_n}\int_0^{\infty}(1-\hat{F}_n(y))dy,$$

where $\bar{x}_n$ is a sample mean $\hat{F}_n$ and is the empirical distribution of claim. Plugging $\hat{F}_{X,n}^s(u)$ into the Pollaczeck-Khinchine formula, we get the estimator for the survival probability [Hipp 1989]

$$\hat{\Phi}(u) = \sum_{n=0}^{\infty}\left(1-\frac{\lambda\mu}{c}\right)\left(\frac{\lambda\mu}{c}\right)^n\left(\hat{F}_{X,n}^s\right)^{*n}(u).$$

If we assumed that λ, μ are known, F is unknown, and the data $X_1, X_2, \ldots, X_n$ are available, then the survival probability is estimated by [Hipp 1989]

$$\hat{\Phi}(u) = \sum_{n=0}^{\infty}\left(1-\frac{\lambda\mu}{c}\right)\left(\frac{\lambda\mu}{c}\right)^n\left(\hat{F}_X^s\right)^{*n}(u),$$

$$\hat{F}_X^s(u) = \frac{1}{\mu}\int_0^{u}(1-F(y))dy,$$

where F is estimated by a discrete signed measure

$$\hat{P}_n(x_i) = \frac{\#\{j:x_j = x_i\}}{n}\left(1-\frac{(x_i-\bar{x}_n)(\bar{x}_n-\mu}{s^2}\right),$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2.$$

When parameter λ and distribution F are unknown, and the observations from a risk process in a time interval $(0,T)$ are available, then the estimator of the survival probability is given by [Hipp 1989]

$$\hat{\Phi}(u) = \sum_{n=0}^{\infty}\left(1-\frac{\hat{\lambda}_T\mu}{c}\right)\left(\frac{\hat{\lambda}_T\mu}{c}\right)^n\left(F_X^s\right)^{*n}(u),$$

where

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 7 (13)

$$F_X^s(u) = \frac{1}{\mu}\int_0^u (1 - \hat{F}_T(y))dy.$$

The intensity λ is estimated by $\hat{\lambda}_T = \frac{N(T)}{T}$, and F is estimated by

$$\hat{F}_T(u) = \frac{1}{N(T)}\sum_{i=1}^{N(T)} 1(X_i \le u).$$

The estimators of the ruin probability found by using the Cramer-Lundberg approximation and the Pollaczeck-Khinchine formula are asymptotically normal. In the paper there will be shown examples illustrating the finite sample behaviour of estimators presented above.

References

Babier J., Beda C., *Approximations of ruin probability by diatomic or diexponential claims*. ASTIN BULLETIN 1992, vol. 22, no. 2, pp. 235-246.

Csörgő S., Teugels J.L., *Empirical Laplace transform and approximation of compound distribution*. Journal of Applied Probability 1990, vol. 27, no. 1, pp. 88-101.

Grandell J., *Aspect of Risk Theory*, Springer-Verglag, New York 1991.

Hipp C., *Estimators and bootstrap confidence intervals for ruin probabilities*, ASTIN BULLETIN 1989, vol. 19, no. 1, pp. 57-70.

CAUSAL INFERENCE USING THE PROPENSITY SCORE

Karl-Heinz Schild, Helmut Sitter
(Department of Business Administration and Economics University of Marburg)

The term 'propensity score' refers to a generic approach for causal inference in observational studies, which was introduced by Rosenbaum and Rubin in [Rosenbaum, Rubin 1983] and has since then been widely used. The basic framework for an application of the propensity score can be described as follows: Given a random sample from a certain population, the following data are observed for each individual i in the sample: A set of (pre-treatment) covariates $\boldsymbol{X}_i$, the treatment (assignment) Z_i and the outcome Y_i, which is the variable of primary interest. In their genuine form, propensity score applications require Z_i to be a binary variable ($Z_i \in \{0,1\}$), where $Z_i = 1$ indicates that the individual received treatment, while $Z_i = 0$ indi-

cates that the individual was in the control group. In the binary treatment case, the propensity score is defined as the probability of receiving treatment given the covariates:

$$p(x) = P(Z_i = 1 | X_i = x).$$

It is, of course, possible to generalize the propensity score to situations where Z_i is a random variable taking values in an arbitrary space Z, for example as $r(z,x) = P(Z = z \mid \boldsymbol{X} = x)$, see e.g. [Imbens 2000; Imai, Van Dyke 2004].

The framework for causal interpretations is the *potential outcome model*. The idea is to stipulate existence of an outcome Y_i^z for all $z \in Z$. The observed (factual) outcome is $Y_i = Y_i^{Z_i}$, while Y_i^z for $z \neq Z_i$ is understood to be the (counterfactual) outcome, had the individual received treatment z instead. For a binary treatment, the causal effect of the treatment on individual *i* is $Y_i^1 - Y_i^0$, and the *average treatment effect* is ATE = E[$Y^1 - Y^0$].

The key assumption underlying the propensity score based treatment evaluation is the conditional unconfoundedness assumption (CUA) (also termed "ignorability of treatment assignment"): Conditional on the covariates X_i, the treatment assignment Z_i and the potential outcomes Y_i^z are independent:

$$Y_i^z \perp Z_i \mid \boldsymbol{X}_i \text{ for all } z \in \mathsf{Z}.$$

This condition means that the pre-treatment covariates must be informative enough to resolve the "confoundedness" or "selectivity" in the treatment assignment process up to the situation of a randomized experiment.

If the CUA holds, it also holds that the Y^z are independent from treatment Z given $p(\boldsymbol{X})$: $Y^z \perp Z \mid \boldsymbol{X} \Rightarrow Y^z \perp Z \mid p(\boldsymbol{X})$. In conjunction with the "balancing property" of $p(\boldsymbol{X})$, namely $\boldsymbol{X} \perp Z \mid p(\boldsymbol{X})$ (which holds irrespective of the CUA), the propensity score is seen to be an ideal tool for matching and stratification: On the strata of constant propensity score, all covariates have the same distribution in the treated and untreated group, while the difference in the average outcome between the groups represents the causal effect of the treatment on the stratum.

Besides the application as a matching tool, one can also make use of the propensity score as a control function in the linear regression of Y on Z. Assume $Y^k = \mu^k + \boldsymbol{X}'\beta^k + U^k$ $(k = 0,1)$, where $E[U^k \mid \boldsymbol{X}] = 0$ and (for simplicity) $U^0 = U^1$, so that ATE $= \mu^1 - \mu^0 + (\beta^1 - \beta^0)' E[\boldsymbol{X}]$. Writing $Y = Y^0 + Z(Y^1 - Y^0)$, we have (provided the CUA holds):

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 7 (13)

$$E[Y|\boldsymbol{X},Z] = E[Y^0|\boldsymbol{X},Z] + E[Z(Y^1 - Y^0)|\boldsymbol{X},Z] = \\ = E[Y^0|\boldsymbol{X}] + Z \cdot E[Y^1 - Y^0|\boldsymbol{X}].$$

Thus, the linear regression of Y on 1, Z and $\boldsymbol{X}$ will yield the ATE as the coefficient of Z. Using the unconfoundedness w.r.t. $p(\boldsymbol{X})$, the same result obtains, if instead of $\boldsymbol{X}$ we condition on $p(\boldsymbol{X})$:

$$E[Y|p(\boldsymbol{X}), Z] = E[Y^0|p(\boldsymbol{X}), Z] + E[Z(Y^1 - Y^0)|p(\boldsymbol{X}), Z] = \\ = E[Y^0|p(\boldsymbol{X})] + Z \cdot E[Y^1 - Y^0|p(\boldsymbol{X})].$$

Thus, provided $E[Y^k|p(\boldsymbol{X})]$ is linear in $p(\boldsymbol{X})$, the regression of Y on 1, Z and $p(\boldsymbol{X})$ will also yield the causal ATE. Analogous results are obtained if a subset of (functions of) the covariates is added to the regressors, as long as the propensity score is included as a control function.

Application: Survival analysis with varying treatment initiation

The objective now is to apply the propensity score to the following problem arising in survival analysis: For a random sample of individuals with covariates $\boldsymbol{X}_i$, the survival time T_i of each individual is observed, if not right-censored by an independent censoring mechanism, in which case the censoring time is observed. Each individual has the chance to begin a treatment at any time $s \geq 0$.. The treatment can be initiated only once, and at most one treatment initiation time $s = S_i$ is observed for each individual *i*. Once received, the treatment has an immediate and persisting effect: It reduces (or enlarges) the individual's hazard rate by the factor e^{α} for the reminder of its life. To model this, a proportional hazard rate model is stipulated that describes the hazard rate h_i of individual *i*, if treatment starts at time *s*:

$$h_i(t|s) = h(t,s|\boldsymbol{X}_i) = e^{\boldsymbol{\beta}'\boldsymbol{X}_i + \alpha 1[t \geq s]} \cdot h_{0(t)} = \begin{cases} e^{\boldsymbol{\beta}'\boldsymbol{X}_i} h_{0(t)}, & \text{if } t < s \\ e^{\alpha} e^{\boldsymbol{\beta}'\boldsymbol{X}_i} h_{0(t)}, & \text{if } t \geq s \end{cases}, \quad (1)$$

where $h_0(t)$ is the baseline hazard. Replacing *s* with the observed initiation time S_i, it is a popular device to estimate such a model with a Cox (partial likelihood) regression [Cox 1972], which involves the time-dependent covariate $x_i(t) = 1[t \geq S_i]$.

Note that eqn. (1) represents the potential outcome model. The observed "treatment assignment" is S_i, and the factual outcome with this treatment assigned is T_i, but eqn. (1) also describes the potential outcome T_i^S, in terms of its hazard rate $h_i(t,s)$, if the assigned treatment were *s*. The conditional unconfoundedness assumption in the present application is:

$S_i \perp T_i^S \mid \boldsymbol{X}_i \forall\, s$. This excludes, for example, the situation that comparable individuals that tend to have an early treatment initiation will have systematically longer survival times anyway, even if they were never treated.

Replicating the argument for the linear regression, the Cox regression for (1) – with s replaced by S_i – can be expected to provide a reliable (consistent) estimate of the causal treatment effect α. This, however, only holds, if the covariates $\boldsymbol{X}$ are exactly those under which the CUA holds; if a component of $\boldsymbol{X}$ is omitted, or is included in the wrong functional form, the estimate of α as the causal treatment effect may become unreliable. We therefore attempt to include the propensity score as a control function in eqn. (1).

To do so, a model for the treatment initiation times S_i is required. Since S_i, like T_i, is a duration, it appears sensible to stipulate a proportional model for the "hazard"-rate $h_i^S(t)$ of S_i:

$$h_i^S(t) = h^S(t, \boldsymbol{X}_i) = e^{\gamma' \boldsymbol{X}_i} h_0^S(t). \tag{2}$$

Again, the intention is to estimate this model by Cox partial likelihood regression. Note, however, that the observation of S_i can be right-censored by T_i (or the censoring time of T_i); this occurs if the individual dies or is censored before treatment is initiated. Since T_i, as opposed to T_i^S, can not be assumed to be independent of S_i, a bias is supposed to occur if the estimation of the Cox model (2) implicitly assumes independence of its censoring. This bias will be ignored in the sequel. Also note that it is possible in (2) (and also in (1)) to have a probability mass at $t = \infty$ – neither the models nor the estimation procedure, the Cox regression, exclude this possibility.

In order to get a clue of how to define the propensity score, we first reformulate the Cox regression (1), (2) by dividing the time axis into episodes $I_j = [t_{j-1}, t_j)$ and using the binary variables

- $Y_{i,j}$ to indicate the event that individual i was dead at t_j given that it was alive at t_{j-1}.
- $\widetilde{Z}_{i,j}$ to indicate the event that individual i was treated at t_j given that it was untreated at t_{j-1}.

We also use the binary variable $Z_{i,j}$ to indicate treatment in I_j (unconditionally). The main result is: The estimation of the Cox partial likelihood regressions produces approximately the same results as estimating the binary response models (by ML assuming independent observations)

$$P(Y_{i,j} = 1 \mid \boldsymbol{X}_i, Z_{i,j}) = G(\tau_j + \boldsymbol{\beta}' \boldsymbol{X}_i + \alpha Z_{i,j}), \tag{3}$$

$$P(\widetilde{Z}_{i,j} = 1 | \boldsymbol{X}_i) = G(\eta_j + \boldsymbol{\gamma}' \boldsymbol{X}_i), \tag{4}$$

where G is a cdf. and τ_j, η_j are interval-specific constants that capture the effect of the baseline hazards $h_0^T(t), h_0^S(t)$. (The precise formulas, although not relevant for the sequel, are: $G(z) = 1 - \exp(\exp(-z))$, $\tau_i = \ln \int_{t_{j-1}}^{t_j} h_0^T(u)du$, $\eta_i = \ln \int_{t_{j-1}}^{t_j} h_0^S(u)du$, see e.g. [D'Agostino et al. 1990] for details.

An important point about these equations is that the $Y_{i,j}$ and $Y_{i',j'}$ as well as the $\widetilde{Z}_{i,j}, \widetilde{Z}_{i',j'}$, can be assumed to be independent from each other except for $i' = i$, $j' = j$. The treatment indicator $Z_{i,j}$ can be expressed as a function of the past $\hat{Z}_{i,j}, k \le j$.

Consider the models (3), (4) for a fixed j, i.e. on the fixed time-interval I_j. Keeping j fixed and letting i run through the risk set for I_j, we are in a standard propensity score setting with $Z_{i,j}$ as the binary treatment variable and $Y_{i,j}$ as the outcome variable (which happens to be binary, too). Denote by p_j the propensity score for the j-th submodel:

$$p_j(\boldsymbol{X}_i) = P) Z_{i,j} = 1 | \boldsymbol{X}_i) = P(S_i < t_j | \boldsymbol{X}_i).$$

Using the the "treatment initiation function" of (2), $F_i^S(t) = F^S(t, \boldsymbol{X}_i) = P(S_i \ge t | \boldsymbol{X}_i)$, we obtain a propensity score in the variable $\boldsymbol{X}$ at each fixed point in time t:

$$p(t, \boldsymbol{X}) = 1 - F^S(t, \boldsymbol{X}). \tag{5}$$

Under the proportional hazard assumption, we have $F^S(t, \boldsymbol{X}) = \left(F_0^S(t)\right)^{\exp(\boldsymbol{\gamma}' \boldsymbol{X})}$. This function is readily estimated as a by-product of the Cox-Regression for (2) by most survival analysis software packages. Thus, we obtain the estimated propensity score $\hat{p}(t, \boldsymbol{X}_i)$ for (3) at any time t (=t_j) at almost no cost.

Having obtained the estimated propensity score for each t, we can use it as a control function in the model for $Y_{i,j}$ in the outcome model (3). Setting $P_{i,j} := \hat{p}(t_j, \boldsymbol{X}_i)$, we can expect a reliable estimation of the treatment effect α from using only the constant (which is here the time dummy τ_j), the treatment indicator $Z_{i,j}$ and the propensity score $P_{i,j}$ as covariates in the the binary response model

$$P(Y_{i,j} = 1 | \boldsymbol{X}_i) = G(\tau_j + \alpha Z_{i,j} + \delta P_{i,j}). \tag{6}$$

However, we do not really need to estimate this binary response model for the $Y_{i,j}$, because we can retranslate it into a Cox model for the proportional hazard rates $h(t \mid S_i, \boldsymbol{X}_i)$ of the survival times T_i, yielding

$$h(t \mid S_i, \boldsymbol{X}_i) = e^{\alpha 1[t \geq s_i] + \delta \hat{p}(t, \boldsymbol{X}_i)} \cdot h_0(t). \tag{7}$$

The preceeding considerations suggest that the Cox regression for this model will produce a reliable estimate of the treatment effect α.

To summarize, the whole procedure consists of two steps:

(1) Estimate a Cox regression for the treatment initiation model (2) using the full set of covariates $\boldsymbol{X}$. Obtain the propensity score $\hat{p}(t, \boldsymbol{X})$, as $1 - F^S(t, \boldsymbol{X})$, where F^S is the "treatment initiation function" (the "survival" function of the treatment initiation model).

(2) Run a Cox regression for the outcome model using (at a minimum) the treatment indicator $1[t \geq S_i]$ and the estimated propensity score $\hat{p}(t, \boldsymbol{X}_i)$, as time-dependent regressors.

Note that the Cox regression in the second step involves *two time-dependent* covariates.

Some of the flexibility of the two-step procedure arises from the fact that we do not have to confine ourselves to just the two time-dependent covariates in the outcome model. The same reliability concerning the estimation of the causal treatment effect α is expected to occur in *any* model of the form

$$h(t, s \mid \boldsymbol{X}_i) = e^{\tilde{\boldsymbol{\beta}} \tilde{\boldsymbol{X}}_i + \alpha 1[t \geq s] + \delta \hat{p}(t, \boldsymbol{X}_i)} \cdot h_0(t), \tag{8}$$

where $\tilde{\boldsymbol{X}}$ consists of a subset and/or functions of the variables in $\boldsymbol{X}$. The crucial requirement is that, in addition to the treatment indicator $1[t \geq S_i]$, the propensity score $\hat{p}(t, \boldsymbol{X}_i)$, is included in the regression. One can, for example, use this feature to check the validity of the propensity score estimate and/or the assumptions (for example CUA): If a variation of the regressors $\tilde{\boldsymbol{X}}$ in (8) produces very different estimates of α, then presumably either one of the (general or parametric) assumptions is violated or the propensity score estimate is not valid. In the ladder case the most likely reason is that the "false covariates" $\boldsymbol{X}$ are included in the treatment initialization model.

Remark: The whole derivation crucially hinges on the assumed independence of the $Y_{i,j}$ (and $\tilde{Z}_{i,j}$). The argument for this is the same sort of approximation that is performed in going from full to partial likelihood in the Cox regression. The main idea behind the Cox partial likelihood regression is to consider – at certain times t_{j-1} – the risk set $\{T_i \geq t_{j-1}\}$ con-

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

sisting of individuals alive at that time. The partial likelihood for the interval I_j is then formed as if the risk set were a random draw from the population. Therefore, treating the $Y_{i,j}$ as independent variables in (3) corresponds to using the Cox partial likelihood method in (1). It is also for this reason that estimating the Cox regression produces approximately the same results as estimating the binary response model (3) by maximum likelihood assuming independent observations.

Results of a simulation: We generate N independent realizations of

- covariates $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$ where X_1 is binary with $P(X_1 = 1) = 0.5, X_2$ is binary with $P(X_2 = 1) = 0.75$, X_3 is uniformally distributed on [–1, 1] and X_4 has a standard normal distribution.
- Weibull-distributed (potential) survival times T^0 with shape parameter $\kappa = 3$ and scale parameter $(1/\lambda)^{1/\kappa}$ where $\lambda = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + + \beta_3 X_3 + \beta_4 X_4$, and

$$\beta_0 = -12, \beta_1 = 0.5, \beta_2 = -0.5, \beta_3 = 0.5, \beta_4 = 0.5.$$

- Weibull-distributed treatment initiation times S with shape parameter $\kappa = 2$ and scale parameter $(1/\lambda)^{1/\kappa}$ where $\lambda = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + + \gamma_3 X_3 + \gamma_4 X_4$ and

$$\gamma_0 = -8.5, \gamma_1 = 0.3, \gamma_2 = -0.3, \gamma_3 = -0.4, \gamma_4 = -0.8.$$

If $S_i > T_i^0$, then T_i is set to T_i^0 and S_i is censored at T_i^0. Otherwise the observed survival time T_i is set to

$$T_i = \sqrt[\kappa]{e^{-\alpha}(T_i^0)^\kappa + (1 - e^{-\alpha})(S_i)^\kappa},$$

with $\alpha = 0.5$ and $\kappa = 3$. The resulting duration times T_i and S_i are conform to the proportional hazards models (1), (2).

The following diagram displays the distribution of the estimated α resulting from of a series of 1000 simulations each using $N = 200$ independent observations.

The first model M0, specifies the outcome equation with a time-independent treatment dummy (and all four covariates). Each of the following five outcome models contain the time-dependent treatment dummy; each of these models is estimated in two variants: without ("-") and with ("+") the estimated propensity score from the initiation model as a control variable. Model M1: No other covariates, Model M2: Only X_3, Model M3: X_1 and X_3 Model M4: X_1, X_2, X_3^2, X_4^2, Model M5: All covariates in correct functional form.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

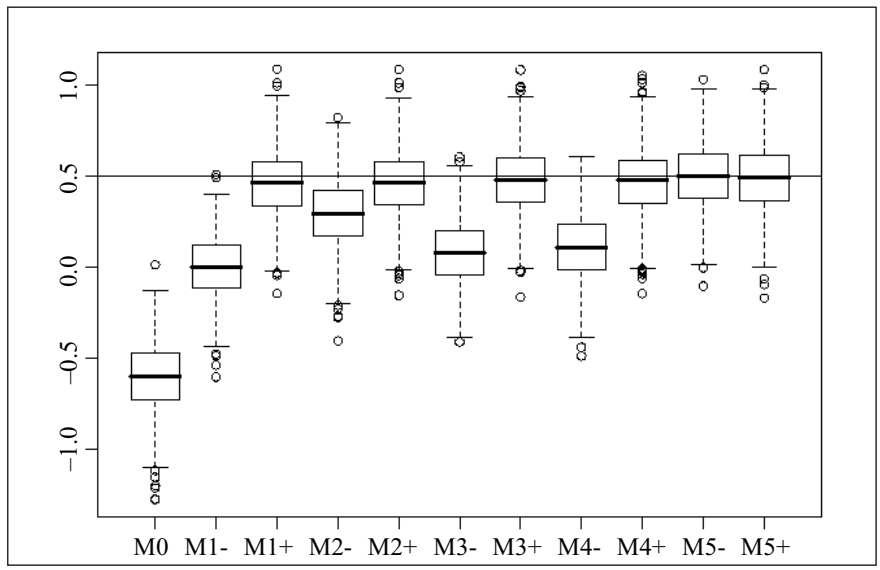

Fig. 1. Diagram

Source: own elaboration.

The diagram shows that a false specification of the outcome model may lead to very biased estimates of the treatment effect α. Including the propensity score as a time-dependent covariate largely alleviates this problem, resulting in estimates which are at least approximately correct. The bias still existing in these estimates is partly due to the non-random censoring of S_i in (2).

References

D'Agostino R., Lee M.L., Belanger A.J., *Relation of pooled logistic regression to time dependent Cox Regression analysis*. Stat. Med., 9 (1990), pp. 1501-1515.

Cox D.R., *Regression models and life tables*. Journal of the Royal Statistical Society Series B, 34 (1972), pp. 187-200.

Imai K., Van Dyke D.A., *Causal inference with general treatment regimes: Generalizing the propensity score*. JASA, 99/467 (2004), pp. 854-866.

Imbens G.W., *The role of the propensity score in estimating dose-response functions*. Biometrika, 87 (2000), pp. 706-710.

Rosenbaum P.R., Rubin D.B., *The central role of the propensity score in observational studies for causal effects*. Biometrika, 70 (1983), pp. 41-55.

ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY

Nr 7 (13)

ON THE DETECTION OF CHANGES IN LINEAR MODELS WITH DEPENDENT ERRORS

Alexander Schmitz (University of Cologne)

1. The monitoring procedure

One of the most central questions in time series analysis is structural stability of the data generating process: How long is a model, previously estimated from a historical period, relevant for steadily arriving data? The sequential test by Chu et al. [1996] yields a procedure to monitor the parameter stability of a linear regression model:

$$y_t = x_t^T \beta_t + \varepsilon_t, \qquad t = 0,1, \ldots, \tag{1}$$

where $x_t = (1, x_{2t}, \ldots, x_{pt})^T$ is the normed p-dimensional regressor array and $\beta_t = (\beta_{1t}, \ldots, \beta_{pt})^T$ is the regression parameter. It is assumed that the parameter is fixed to a certain initial value β_0 on the historical period of length m, the so-called "noncontamination" assumption:

$$\beta_t = \beta_0, \qquad t = 0,1, \ldots, m \tag{2}$$

After the arrival of each new data from time $m+1$ onwards, we are interested in the "no change" null hypothesis $\mathrm{H}_0: \beta_t = \beta_0$, $t = m+1, \ldots$, versus the "one-time parameter shift" alternative H_A that the parameter shifts from β_0 to β_* at some time $m + k^*$. In this setting the change-point k^*, the initial value β_0 and the parameter shift $\beta_* - \beta_0$ are assumed to be unknown. The monitoring procedure of Chu et al. [1996] is designed to stop monitoring and reject the null hypothesis at time $\tau(m)$, according to the first excess of a detector $\hat{Q}_m(\cdot)$ over a boundary function $g_m^*(\cdot)$, i.e. $\tau(m) = \inf\left\{k: \left|\hat{Q}_m(k)\right| > \sigma c(\alpha) g_m^*(k)\right\}$, where σ is a positive constant and $c(\alpha)$ is a critical constant. Moreover, we set $\varnothing = \infty$, if the path of the detector never exits the boundary.

The task is to determine the detector, the boundary function and the critical constant, such that the false alarm rate is asymptotically fixed to a prescribed level and that the power of the testing procedure is asymptotically one, i.e.

$$\lim_{m\to\infty} P\left(\tau(m) < \infty \mid H_0\right) = \alpha \quad \text{and} \quad \lim_{m\to\infty} P\left(\tau(m) < \infty \mid H_A\right) = 1.$$

Consider first the detector. Denote

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

$$\hat{\beta}_m = \left(\sum_{i=1}^{m} x_i x_i^T \right)^{-1} \sum_{i=1}^{m} x_i y_i \quad \text{and} \quad \hat{\varepsilon}_i = y_i - x_i^T \hat{\beta}_m^T$$

the least squares estimator for β_0, relying only on the historical period, and the *i*-th empirical residual. Following Chu et al. [1996], we employ a cumulated sum type detector (CUSUM):

$$\hat{Q}_m(k) = \sum_{i=m+1}^{m+k} \hat{\varepsilon}_i, \quad k = 1, 2, \ldots \tag{3}$$

Horváth et al. [2004] introduced a class of boundary functions being analytical convenient for the CUSUM monitoring. Therefore we choose

$$g_m^*(k) = m^{1/2} \left(1 + \frac{k}{m}\right) \left(\frac{k}{m+k}\right)^{\gamma}, \quad 0 \le \gamma \le 1/2. \tag{4}$$

The parameter γ is a so-called tuning constant influencing the detection ability.

2. Model assumptions and results

In contrast to Horváth et al. [2004] who assumed independent errors, we allow a certain dependency among the error sequence instead. This is a common approach within the econometric change-point literature, cf. Perron and Qu [2007]. Following the α-mixing concept, the measure of dependence between two σ-fields G and H is given by

$$\alpha(G,H) = \sup\left\{ \left| P(A \cap B) - P(A)P(B) \right| : A \in G, B \in H \right\}.$$

Let F_k^l denote the σ-field generated by the set of errors $\{\varepsilon_k, \ldots, \varepsilon_l\}$, where k and l are consecutive integers. We assume a strong mixing error sequence, i.e.

$$\alpha(n) = \sup_{1 \le p \le \infty} \alpha\{F_1^p, F_{p+n}^{\infty}\} \to 0 \quad (n \to \infty). \tag{5}$$

This property indicates that the present innovations are asymptotical independent from the far distant future innovations.

Next, we assume that the error sequence obeys a uniform weak invariance principle. Let σ and δ be positive constants and let $\left\{ W_{i,m}(t) \right\}_{0 \le t < \infty}$ denote a standard Wiener process, for each $i = 0,1$, and for all $m = 1,2, \ldots$, such that a weighted approximation on the historical period and on the monitoring sequence holds:

$$\sup_{1\le k\le m} k^{-1/(2+\delta)}\left|\sum_{i=1}^{k}\varepsilon_i-\sigma W_{0,m}(k)\right|=O_p(1)\quad(m\to\infty). \tag{6}$$

$$\sup_{1\le k\le m} k^{-1/(2+\delta)}\left|\sum_{i=m+1}^{m+k}\varepsilon_i-\sigma W_{1,m}(k)\right|=O_p(1)\quad(m\to\infty). \tag{7}$$

This kind of approximations can be derived by imposing certain moment conditions and a accurate rate of decay of the mixing coefficient $\alpha(n)$, cf. e.g. assumption A5 in Perron and Qu [2007].

We also need some regularity conditions on the stochastic regressor sequence. The Euclidean norm of vectors and matrices are denoted by $\|\cdot\|$. We assume that:

$$\left\|\frac{1}{m}\sum_{i=1}^{m}x_i x_i^T-C\right\|=O_{a.s.}(m^{-\tau})\quad(m\to\infty) \tag{8}$$

holds for some $\tau>0$ and for some positive definite matrix C. Basically, this assumption rules out trending regressors. And we assume further:

$$\left\|\sum_{i=1}^{m}x_i\varepsilon_i\right\|=O_P(m^{1/2})\quad(m\to\infty). \tag{9}$$

The last assumption is a technical condition in order to permit regressors not necessarily independent of the errors.

We state our main result: Under the null hypothesis, suppose (1)-(9) hold, then we have:

$$\lim_{m\to\infty}P\left(\frac{1}{\sigma}\sup_{1\le k<\infty}\frac{|\hat{Q}_m(k)|}{g_m^*(k)}>c\right)=P\left(\sup_{0<t\le 1}\frac{|W(t)|}{t^{\gamma}}>c\right).$$

The limit distribution is a functional of the Wiener process. Selected quantiles are given in Horváth et al. [2004]. An application of the monitoring procedure in practice requires a consistent estimation of the unknown parameter σ. As a consequence of using invariance principles for dependent random variables σ^2 is the long run variance, i.e.

$$0<\sigma^2=E\varepsilon_1^2+2\sum_{k=2}^{\infty}E\varepsilon_1\varepsilon_k<\infty.$$

Consistent estimators are available using a “non-overlapping blocking” approach. Moreover, under the alternative, asymptotic power one of

the monitoring procedure can be shown, if the mean regressor is not orthogonal to the parameter shift $\beta_* - \beta_0$. For details and proofs we confer to Schmitz and Steinebach [2008].

3. Remarks

For an adequate monitoring of econometric data, it seems reasonable to choose time series model in order to form the errors of the linear regression model. The crucial step in our framework is to ensure the strong mixing property. The simplest model but still useful in the macroeconomic context is the AR(1) (autoregressive of order one) model, cf. e.g. Hansen [2001]. It is given by the recurrence equation: $X_t = \phi X_{t-1} + Z_t$, where $\phi \in (0,1)$ and $\{Z_t\}_{t=1,2,\ldots}$ are uncorrelated noise variables. We point out that in the case of discrete noise variables the mixing property can be violated, cf. Andrews [1984]. In the case of independent and identically distributed noise variables with an absolute continuous probability distribution, Athreya and Pantula [1986] provide conditions on the probability density to ensure the strong mixing property. The classical ARCH(p) (autoregressive conditionally heteroskedastic of order p) model of Engle [1982] is defined by the equations

$$X_t = \sigma_t Z_t \quad \text{and} \quad \sigma_t^2 = c_0 + \sum_{i=1}^{p} b_i X_{t-i}^2,$$

where all constants are positive. This model and its generalisations play a central role in finance and econometrics, cf. e.g. Aue et al. [2006]. We confer to Carrasco and Chen [2000] for conditions to ensure the strong mixing property of ARCH(p) models and its generalisations.

References

Athreya K.B., Pantula S.G., *A note on strong mixing of ARMA processes*. Statistics and Probability Letters, 4 (1986), pp. 187-190.

Andrews D.W.K., *Non-strong mixing autoregressive processes*. Journal of Applied Probability, 21 (1984), pp. 930-934.

Aue A., Horváth L., Hušková M., Kokoszka P., *Change-point monitoring in linear models*. Econometrics Journal, 9 (2006), pp. 373-403.

Carrasco M., Chen X., *β-mixing and moment properties of RCA models with application to GARCH(p,q)*. Comptes Rendus de l'Académie des Sciences Paris 331 Série I (2000), pp. 85-90.

Chu C.S.J., Stinchcombe M., White H., *Monitoring structural change*. Econometrica, 64 (1996), pp. 1045-1065.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

Engle R.F., *Autoregressive conditional heteroskedasticity with estimates of the variance of the United Kingdom inflation*. Econometrica, 50 (1982), pp. 987-1007.

Hansen B.E., *The new econometrics of structural change: Dating changes in U.S. labor productivity*. Journal of Economic Perspectives, 15 (2001), pp. 117-128.

Horváth L., Hušková M., Kokoszka P., Steinebach J.G., *Monitoring changes in linear models*. Journal of Statistical Planning and Inference, 126 (2004), pp. 225-251.

Perron P., Qu Z., *Estimating and testing structural changes in multivariate regressions*. Econometrica, 75 (2007), pp. 459-502.

Schmitz A., Steinebach J.G., *A note on the monitoring of changes in linear models with dependent errors*, Preprint University of Cologne (2008), pp. 1-14.

ON HEALTH-ORIENTED LIFESTYLE RESEARCH

Cyprian Kozyra (Wroclaw University of Economics)

Student Scientific Circle on Survey Research conducted under supervision of author of this paper research on health-oriented lifestyle in year 2007. Main aims of the survey research were: investigating of respondents' awareness of influence of lifestyle on health and assessment of lifestyle practiced by respondents. Population of concern were students of Wroclaw University of Economics. We defined lifestyle as all behaviors chosen by respondents in aim to mould their own life. The main limitation of such survey is investigating only respondents' perceptions without access to real data about their health and lifestyle.

These detailed questions were investigated in research: What do respondents mean about healthy lifestyle? Are they aware of influence of lifestyle on health? How do they assess their lifestyle and why? Do they feel, that their lifestyle should change and how to do it? What is the main respondents' goal of practicing healthy lifestyle? These questions were employed in wording questionnaire consisting twelve (some of them were very complex) items, on which answers are presented in paper.

Research sample was designed by means of random cluster sampling, because lack of available list of all units of population of concern. Additionally during random selection we decided to apply two-way stratification according to faculty (3 Wroclaw faculties of university) and year of study (5 years). Only one cluster from all fifteen strata was selected, but we did not collected data from six student groups, so, because of these missing data, only stratification according to faculties was applied in analysis. Using complex samples methodology is more difficult than simple ran-

dom sampling, because computation of unbiased estimates and their variances is not available in standard statistical programs. For instance unbiased estimates of frequency in case of stratified cluster sampling is (see [Miszczak 2004, p. 95]): $\hat{p}_{wg} = \sum_{h=1}^{L} \frac{W_h}{m_h} \sum_{k=1}^{m_h} \hat{p}_{hk}$, where $\hat{p}_{hk}$ is estimator of frequency in group k of stratum h, m_h is number of groups in stratum h, and W_h is share of stratum h in whole population. One can notice that unbiased estimation for equal-numerous cluster could be accomplished in standard statistical software by means of special weighting units according to strata.

Additionally for complex samples methods of statistical inference should be modified, e.g. test to verify independence in contingency table should be (see [Bracha 1998, p. 149]) conducted not only using well-known Pearson statistics $\chi^2_P = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(p_{ij} - p_{i+} p_{+j})^2}{p_{i+} p_{+j}}$, but also variances of frequency estimators should be compared with estimators to calculate modified chi-square statistics $\chi^2_m = \frac{\chi^2_m}{\lambda_0}$, where $\lambda_0 = \frac{n}{rc-1} \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{V(p_{ij})}{p_{ij}}$. Variances of estimators are not calculated in standard statistical software and in this research their calculations are questionable because of missing data.

Main result from survey are as follows: almost all respondents agree (76% strongly agree and 23% rather agree) that generally lifestyle has influence on health, almost all respondents (young students) perceive their health as good (21% very good and 73% rather good), almost all respondents agree that their lifestyle has influence on their own health (49% strongly agree and 45% agree, notice difference with answers on questions about general influence of lifestyle on health). Self-assessment of own lifestyle is not so unambiguous – most of respondents assess their lifestyle as rather good for health (65%), but many (23%) assess as neither good nor bad, only 6% asses as very good for health. Most of respondents (54%) report that they did not change their lifestyle during last year, 28% changed rather for the better and 12% changed rather for the worse. Most of respondents (31% strongly agree and 50% rather agree) are willing to change their lifestyle for the better.

Main reasons for improving lifestyle are: will to improve mental and physical state (73% respondents agree with it), influence of close people (47%) and deteriorating health condition (41%). These items were not

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 7 (13)

often selected by respondents as reasons for improving lifestyle: healthy oriented campaign (only 4%) and fashion (18%). Main reasons for not improving lifestyle are: lack of time (64%) and laziness (58%), but lack of money is not the reason (19%). Respondents assess some aspects of their lifestyle on scale 1-5 (1 is the worst for health, 5 is the best) and average answers are as follows: practicing sport – 3.00, healthy eating – 3.16, suitable clothing – 3.38, fighting against addictions – 3.26, medical control tests – 2.44, taking medication – 2.99, vitamin supplements – 2.92, avoiding risky sexual behavior – 3.54. The best assessments are thus for avoiding risky sexual behavior and the worst for medical control tests.

Dependencies between all response categories regarding main questions were tested by means of chi-square statistics. Significant dependencies are as follow: there is dependence between general influence of lifestyle on health and influence of own lifestyle on health, dependence between overall influence of lifestyle and change of lifestyle in last year, dependence between self-assessment of health and of lifestyle, dependence between influence of own lifestyle on health and change of lifestyle in last year, dependence between influence of own lifestyle and willingness to change, and dependence between self-assessment of lifestyle and change of lifestyle in last year. All these significant dependencies are positive in the sense, that better answers (from health point of view) of one questions correspond to better answers of second questions.

In conclusion we can say that results of research give positive image, but rather nobody knows whether it is the sign of reality or the sign of respondents' expectations of their life. Complex sample methodology used in this research could be useful in some cases, but more difficult in calculation.

References

Barnett V., *Elementy teorii pobierania prób*, PWN, Warszawa 1982.

Bracha Cz., *Metoda reprezentacyjna w badaniu opinii publicznej i marketingu*, EFEKT, Warszawa 1998.

Miszczak W., *Projektowanie próby*, AE, Wrocław 2004.

Som R.K., *Practical sampling techniques*, 2nd ed., New York Marcel Dekker, New York 1996.

Steczkowski J., *Metoda reprezentacyjna w badaniach ekonomiczno-społecznych*, PWN, Warszawa-Kraków 1995.