ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 206, 2007

*Czesław Domański**

VERIFICATION OF HYPOTHESES CONCERNING PARAMETERS OF THE REGRESSION MODEL FOR COMPLEX SAMPLES

Abstract. The paper considers the linear regression function $\mathbf{y} = \boldsymbol{\beta}\mathbf{x} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}$ is a vector of unknown parameters and $\boldsymbol{\varepsilon}$ is a rest component. In case of complex samples some modifications of test statistics should be made. Results of simulation study revealed that the verification of the hypothesis H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ should be conducted by means of modified test F.

Key words: complex samples, test F, testing, design effect, χ^2 test.

1. INTRODUCTION

One of the fastest developing areas of the statistical methods application is social research. Nowadays, every researcher analyzing results of opinion polls refers to statistical tools, in the form of computer programs most often, in order to sum up and present their results. Researchers believe that results calculated by means of these computer programs and interpreted according to rules they were taught are the basis for drawing conclusions and setting research hypotheses. It happens quite frequently that interpreting results of sample investigations, researchers forget that they are biased with random errors of the sample (they can be biased with non-random errors as well, which is very often caused by researchers themselves). Thus, for example, they interpret differences in proportions of the support for political parties as if all electors and not only the sample were subjected to the investigation. More often we refer to statistical inference methods taking into account dependence analysis. Still, one of the most popular statistical tools among social researchers is χ^2 test which is used for determining

* Professor, Chiar of Statistical Methods, University of Łódź.

statistical significance of relations between qualitative variables presented by means of contingency tables. The expression "relation is statistically significant" is one of the most popular phrases which is used in sociological analyses of investigations' results. Both those who know the concept of confidence intervals or proportions tests, and those who use χ^2 test, are concerned about the fact whether conditions of their applicability are satisfied. However, they seldom realize that popular computer programs calculate standard errors and statistical tests results by the assumption that the sample was sampled by means of the simple sampling scheme. Consequently, we have to do with the rejection of independence hypotheses as well as the inference about the interdependence of investigated variables occurrence when, using proper analysis methods, they could be considered statistically insignificant.

Sociologists involved in designing investigations usually know that their samples are complex and sampled as results of multistage cluster schemes with the use of stratification. Using weighing, at least at the basic stage, is very common especially in research agencies. Users of data sets of popular investigations such as Polish General Opinion Poll or data revealed by the Public Opinion Research Center also encounter weights which compensate for various probabilities of the inclusion in accepted sampling scheme. They include the post-stratification for the sake of a few demographic characteristics (investigations reports seldom mention weighing in order to compensate for the problem of not realizing the measurement of the part of sampled sample). However, even if we use weights properly and decrease sizes of the bias of parameters estimation with the non-random error, we seldom take note of the fact that a weighted sample, even if initially sampled by means of simple sampling, leads to different values of standard error estimates and the change of statistical tests results but at the same time changing the estimators variance. The influence of accepted sampling schemes and the application of clusters, stratification and diverse inclusion probability as well as samples weighing on the statistical inference is sometimes even realized but does not lead to the application of adequate statistical techniques and methods. Up till now, it has been most often a result of limitations of statistical software. Even though there existed such specialized tools as SUDAAN, WesVar, complex samples analysis module in STATA (the program which unfortunately was not very popular in Poland) or even such free tools as CLUSTERS, limited availability and knowledge of this kind of software made using it very difficult. As a result, the belief that "what we commonly do in the area of data analysis we do poorly" has become quite widespread.

2. PROBLEM FORMULATING

In sample surveys the estimation of unknown populations' parameters is conducted in most cases not on the basis of simple samples but complex ones. This observation presumably inclined L. Kish (1965) to introduce the divergence meter, called "design effect", between the simple sample and the complex one. The meter was defined as follows:

$$deff(t) = \frac{D^2(t)}{D^2(y')} \tag{1}$$

where $D^2(y')$ is the variance of the simple mean sample sampled with replacement (lpzz), and $D^2(t)$ – is the variance of the estimator of the mean $\overline{Y}$ of the characteristic Y by the sampling scheme (we assume that statistics $\overline{y}'$ and t were constructed on the basis of samples of the same size n). The value $deff(t)$ enables to compare variances of various estimators of the mean $\overline{Y}$ (for example quotient, regression) constructed on the basis of data samples sampled according to various sampling schemes with the variance of the sample mean for the scheme with replacement. If $deff(t)$ does not differ from 1 much, then the sample can be treated as the simple one and we can apply e.g. classic significance tests in order to verify the hypothesis $H_0: \overline{Y} = \overline{Y}_2$, where $\overline{Y}_0$ is hypothetical value of the mean of the characteristic Y or the hypothesis $H_0: \overline{Y}_1 = \overline{Y}_2$ where $\overline{Y}_1(\overline{Y}_2)$ is the mean of the first (the second) population. If $deff(t) < 1$, then the actual size of the test is smaller than the assumed one. If $deff(t) > 1$, then the actual probability of the type I error is bigger than the assumed significance level. In sample surveys samples are most often sampled according to very complex schemes, for example two-stage sampling with I stage units stratification (jps). And, at the same time, in jps strata they are sampled with probabilities proportional to the value of the particular additional characteristic X and without replacement (lppxbz), while at the second stage we have to do with the simple sampling without replacement (lpbz). Taking such schemes into consideration, $deff(t)$ can even attain **the value 8** (Kish 1965). It turns out that the more homogeneous jps and the bigger the average fraction of the sample at the second stage, the bigger $deff(t)$. Under such circumstances, the real test size can even exceed 0,50 by assumed 0,05 (Bracha (1998). Then, testing the null hypothesis by means of the classic test assuming that the sample is simple and the test statistic modification is unnecessary, becomes pointless. The influence of the sampling effect on the inference based on complex samples is described by the effective sample size defined by L. Kish (1965) by means of the following formula:

$$n_e = \frac{n}{deff(t)} \tag{2}$$

where n is the real sample size.

Cz. Bracha (2003) presented a few estimators *deff*, what enables to carry out the statistical inference according to the assumed confidence coefficient or the significance level.

3. VERIFICATION OF HYPOTHESES CONCERNING PARAMETERS IN THE REGRESSION MODEL

N-element parent population $U = \{1, 2, ..., N\}$ in which we can observe random characteristics $Y, X_1, X_2, ..., X_k$ is given. The characteristic Y will be called the interpreted one whereas characteristics $X_1, X_2, ..., X_k$ – will be called the interpretative one. Values of these characteristics will be denoted by $Y_j, X_j, ..., X_{jk}$.

We will investigate the linear regression function

$$Y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \tag{3}$$

where $\beta_1, ..., \beta_k$ are unknown parameters and ε is the rest component.

The equivalent of the model (3) for the sample is the formula

$$\mathbf{y} = \boldsymbol{\beta}\mathbf{x} + \mathbf{e} \tag{4}$$

where $\mathbf{e}$ is the vector of random components.

In case of the sample chosen according to the simple sampling scheme with replacement as the estimator of the vector $\boldsymbol{\beta}$, it is necessary to assume

$$\mathbf{b} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} \tag{5}$$

Verification of hypotheses concerning the vector $\boldsymbol{\beta}$ is considered assuming that the random component vector $\mathbf{e}$ has n-dimensional normal distribution with the null vector of expected values and covariance matrix Σ dependent on the applied sample sampling scheme. It turns out that despite the finite population, this assumption can be made only if n and N are big enough (Bracha 1998).

Hypotheses that we are interested in, can be described by the following general linear hypothesis (Theil 1979):

$$H_0: \mathbf{A}\boldsymbol{\beta} = \mathbf{d} \quad \text{against} \quad H_1: \mathbf{A}\boldsymbol{\beta} \neq \mathbf{d} \tag{6}$$

where:

$\mathbf{A} = [a_{pi}]_{q \times k}$ is the given matrix of the rank q $(q < k)$,

$\mathbf{d} = [d_p]_{q \times 1}$ is the given vector.

Below we will present two special cases of the hypothesis (6). We obtain the first one assuming that $\mathbf{A} = \mathbf{I}$ and $\mathbf{d} = \boldsymbol{\beta}_0$. Then

$$H_0: \boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \text{against} \quad H_1: \boldsymbol{\beta} \neq \boldsymbol{\beta}_0 \tag{7}$$

In the second case, matrix $\mathbf{A}$ is the first versor (the vector in which 1 comes first and the remaining coordinates equal to 0), whereas the vector $\mathbf{d}$ is reduced to the scalar β_{l0}.

$$H_0: \beta_l = \beta_{l0} \quad \text{against} \quad H_1: \beta_l \neq \beta_{l0} \tag{8}$$

From our deliberations it follows that the form of the test statistic depends on the sample sampling scheme.

If the sample was sampled according to the scheme with replacement, then in order to verify H_0 of the form (6) it is necessary to apply test F based on the form of statistic (R a o 1982)

$$F = \frac{(\mathbf{Ab} - \mathbf{d})^{\mathrm{T}}[\mathbf{A}(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{A}^{\mathrm{T}}]^{-1}(\mathbf{Ab} - \mathbf{d}))/q}{(\mathbf{y} - \mathbf{y})^{\mathrm{T}}(\mathbf{y} - \mathbf{y})/(n - k)} \tag{9}$$

Assuming truthfulness of hypothesis (6) and normality of random component $\mathbf{e} \sim \mathrm{N}(0, \sigma^2\mathrm{I})$, and moreover, if matrix $\mathbf{x}$ is fixed, then the statistic defined by means of formula (9) has F Snedecor's distribution of q and $n - k$ degrees of freedom.

Verifying hypothesis H_0 defined by the formula (7), we obtain a simpler form of statistic (9)

$$F = \frac{\|\mathbf{xb} - \mathbf{x}\boldsymbol{\beta}\|^2/k}{\|\mathbf{y} - \mathbf{xb}\|^2/(n - k)} = \frac{\mathbf{e}^{\mathrm{T}}\mathbf{x}(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}\mathbf{e}/k}{\mathbf{e}^{\mathrm{T}}[I - \mathbf{x}(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}]\mathbf{e}/(n - k)} \tag{10}$$

If the null hypothesis (7) is true, the statistic defined by the formula (10) has the distribution $F(k, n - k)$.

Now we draw our attention to the hypothesis defined by the formula (8). Let us notice that the statistic F defined by the formula (9) takes a simple form

$$F = \frac{(b_i - \beta_{i0})^2}{s^2 z_{11}} \tag{11}$$

where z_{11} is the first diagonal element of the matrix $(\mathbf{x}^T\mathbf{x})^{-1}$.

The statistic defined by the formula (11), assuming truthfulness of the hypothesis of the form (8), has the distribution $F(1, n-k)$. In practice, we frequently verify hypotheses

$$H_0 : \beta_i = \beta_{i0} \quad \text{against} \quad H_1 : \beta_i > \beta_{i0} \quad \text{or} \quad H_1 : \beta_i < \beta_{i0} \tag{12}$$

instead of hypothesis of the form (8).

Hypotheses defined by (12) are verified by means of the test based on the statistic

$$t_i = (b_i - \beta_{i0})/\sqrt{s^2 z_{ii}} \tag{13}$$

which has Student's distribution t of $n-k$ degrees of freedom assuming truthfulness of the hypothesis H_0. If the sample is not sampled according to the scheme with replacement, then the covariance matrix of random components $\mathbf{e}$ is not the scalar one. Consequently, the numerator and the denominator of the formula (10) do not have distributions $\chi^2(k)$ and $\chi^2(n-k)$, adequately (R a o 1982). In case of the complex sample in order to verify the hypothesis (7), one should not use the statistic given by the formula (10).

Let us now consider symmetric and positively defined matrix $\mathbf{V}$. There exists such a non-singular matrix $\mathbf{C}$, for which the following conditions are satisfied:

$$\mathbf{CVC}^T = \mathbf{I} \quad \text{and} \quad \mathbf{C}^T\mathbf{C} = \mathbf{V}^{-1} \tag{14}$$

If the formula (4) is two-sidedly and left-handedly multiplied by $\mathbf{C}$ (G o l d b e r g e r 1972)

$$\mathbf{Cy} = \mathbf{CxB} + \mathbf{Ce} \tag{15}$$

and we accept denotations

$$\dot{\mathbf{y}} = \mathbf{Cy}, \dot{\mathbf{x}} = \mathbf{Cx} \quad \text{and} \quad \dot{\mathbf{e}} = \mathbf{Ce} \tag{16}$$

then, we obtain

$$\dot{\mathbf{y}} = \dot{\mathbf{x}}\boldsymbol{\beta} + \dot{\mathbf{e}} \tag{17}$$

Let us notice that the distribution of random component $\mathbf{e} \sim N(0, S^2\mathbf{V})$ is $\dot{\mathbf{e}} \sim N(0, S^2\mathbf{V})$. The sample $[\dot{\mathbf{y}}:\dot{\mathbf{x}}]$ can be treated as the simple one. If we apply the least squares method to the sample $[\dot{\mathbf{y}}:\dot{\mathbf{x}}]$ then, we obtain the estimator for $\boldsymbol{\beta}$ identical with generalized least squares method applied to the sample $[\mathbf{y}:\mathbf{x}]$. In order to verify the hypothesis (7) we must apply the statistic

$$\dot{F} = \frac{\dot{\mathbf{y}}\mathbf{b} - \dot{\boldsymbol{\beta}}^2/k}{\dot{\mathbf{y}} - \dot{\mathbf{x}}\mathbf{b}^2/(n-k)} = \frac{(\mathbf{b}-\boldsymbol{\beta}^T)\mathbf{x}^T\mathbf{V}^{-1}\mathbf{x}(\mathbf{b}-\boldsymbol{\beta})/k}{(\mathbf{y}-\mathbf{x}\mathbf{b}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{x}\mathbf{b})/(n-k)} =$$

$$= \frac{\mathbf{e}^T\mathbf{V}^{-1}\mathbf{x}(\mathbf{x}^T\mathbf{V}^{-1}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{V}^{-1}\mathbf{e}/k}{\mathbf{e}^T[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{x}(\mathbf{x}^T\mathbf{V}^{-1}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{V}^{-1}]\mathbf{e}/(n-k)} \tag{18}$$

In case of two-stage scheme it would be easier to write the statistic (18) in the following form:

$$\dot{F} = \frac{\dot{\mathbf{e}}^T\mathbf{C}\mathbf{x}(\mathbf{x}^T\mathbf{V}^{-1}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{C}^T\dot{\mathbf{e}}/k}{\dot{\mathbf{e}}^T[\mathbf{I} - \mathbf{C}\mathbf{x}(\mathbf{x}^T\mathbf{V}^{-1}\mathbf{x})\mathbf{x}^T\mathbf{C}^T]\dot{\mathbf{e}}/(n-k)} \tag{19}$$

Matrices

$$\dot{\mathbf{Q}} = \mathbf{C}\mathbf{x}(\mathbf{x}^T\mathbf{V}^{-1}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{C}^T \tag{20}$$

and

$$\dot{\mathbf{T}} = \mathbf{I} - \mathbf{C}\mathbf{x}(\mathbf{x}^T\mathbf{V}^{-1}\mathbf{x})\mathbf{x}^T\mathbf{C}^T \tag{21}$$

are idempotent matrices of ranks k and $n-k$ adequately. What is more, $\dot{\mathbf{Q}}\dot{\mathbf{T}} = 0$, which proves that square forms occurring in the numerator and the denominator of the formula (19) are stochastically independent.

In practice, we do not know the matrix $\mathbf{V}$ and that is why we estimate it basing on the sample. Therefore, in order to verify H_0 from the formula (7) one should use the following statistic:

$$\dot{F} = \frac{(\hat{\mathbf{b}}-\boldsymbol{\beta})^T\mathbf{x}^T\mathbf{V}^{-1}\mathbf{x}(\hat{\mathbf{b}}-\boldsymbol{\beta})/k}{(\mathbf{y}-\mathbf{x}\hat{\mathbf{b}})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{x}\hat{\mathbf{b}})/(n-k)} \tag{22}$$

Assuming truthfulness of H_0, big sample size and fixed matrix $\mathbf{x}$, the statistic (25) has the approximate distribution $F(k, n-k)$.

We can now ask the question whether in case of the two-stage sampling we can apply the following statistic instead of the statistic (22):

$$F(\beta) = \frac{(\hat{\mathbf{b}} - \beta)^T \mathbf{x}^T \mathbf{x} (\hat{\mathbf{b}} - \beta)/k}{(\mathbf{y} - \mathbf{x}\hat{\mathbf{b}})^T (\mathbf{y} - \mathbf{x}\hat{\mathbf{b}})/(n-k)} =$$

$$= \frac{\mathbf{e}^T \mathbf{x} (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{e}/k}{\mathbf{e}^T [\mathbf{I} - \mathbf{x}(\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T] \mathbf{e}/(n-k)} \tag{23}$$

The formula (23) is an equivalent of the statistic from the formula (10), except that b was replaced with b.

C. F. Wu, D. Holt and D. J. Holmes (1988) suggested another modification of the test statistic defined by the formula (9). If n-element sample is sampled according to the scheme with replacement, then $D^2(b) = \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}$ proceeds, assuming that the matrix $\mathbf{x}$ is fixed, while $\mathbf{b}$ is defined by the formula (5). For other sampling schemes, we have

$$D^2(\mathbf{b}) = \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}(\mathbf{x}^T\mathbf{V}\mathbf{x})(\mathbf{x}^T\mathbf{x})^{-1} = \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{D} \tag{24}$$

where

$$\mathbf{D} = (\mathbf{x}^T\mathbf{V}\mathbf{x})(\mathbf{x}^T\mathbf{x})^{-1} \tag{25}$$

The matrix $\mathbf{D}$ given by the formula (25) is called the misspecification effect matrix (meff) (Scott, Holt 1982). Assuming that at least one out of two matrices $\mathbf{x}^T\mathbf{x}$ or $\mathbf{D}$ is the diagonal matrix, C. F. Wu, D. Holt and D. J. Holmes (1988) applied the following statistic to verify the hypothesis (7)

$$F_m = F \Big/ \left\{ \frac{\mathrm{tr}(\mathbf{QV})/k}{[n - \mathrm{tr}(\mathbf{QV})]/(n-k)} \right\} \tag{26}$$

$$\mathrm{tr}(\mathbf{QV}) = \mathrm{tr}(\mathbf{D}) = \sum_{l=1}^{k} D^2(b_l)/D^2(b_l|\rho = 0) \tag{27}$$

while

$$\mathbf{Q} = \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T$$

where F is defined by the formula (10) and the matrix $\mathbf{V}$ is defined by the formula (14).

When the null hypothesis is true and the sample is big enough, the statistic (26) has, approximately, the distribution $F(k, n-k)$. If the matrix $\mathbf{V}$ is the unitary one, as it is in case of the scheme with replacement, then, the matrix $\mathbf{D}$ from the formula (25) is also the unitary one and the statistic F_m is identical with the statistic F.

In case of complex schemes (for example two-stage), the matrix $\mathbf{V}$ is unknown and therefore it must be estimated on the basis of the sample. To this end, one should use the two-stage procedure. If we do not want

to improve the efficiency of the vector β estimation by means of the application of generalized least squares method instead of least square method, then applying test statistic defined by the formula (25) does not seem to be less laborious than using the statistic $\dot{F}$ from the formula (19). It can be concluded that the estimation of the matrix $\mathbf{V}$ is necessary in both cases. Moreover, in the second case there is no limit when it comes to the form of the matrix $\mathbf{x}^T\mathbf{x}$ (it is diagonal when, for example, we consider the model of the variance with the single classification analysis).

4. FINAL REMARKS

Presented deliberations suggest that verifying the described hypotheses one should take into consideration the fact that the sample is complex. However, from the presented formulas it does not result explicitly that including particular sampling scheme improves the test size significantly. That is why, in order to get at least approximate answer to the problem, simulative investigation based on three-dimensional populations was conducted. The investigation generated five 1000-element populations. In two-stage scheme units size of the first and the second degrees ($m = 10, 15, 20$) were differentiated. Every considered variant in experiment was repeated 300 times for sample size $n = 50, 100, 150$ (see Tab. 1). The investigation includes also, except for two-stage sampling scheme, simple sampling elements as well as sampling without replacement. The investigation gives good grounds to state explicitly that verifying parameters of regression model on the basis of complex samples it is necessary to apply modified tests including the so called sampling scheme effect ($deff$(t)).

Table 1

Size of the test F for $\alpha = 0.05$

n	Statistics F	
	F	F_m
50	0.281	0.066
100	0.192	0.061
150	0.168	0.051
200	0.126	0.049

Source: own calculations.

In particular, the size of test F for degree $\alpha = 0.05$, in case of applying the classic form of the statistic, exceeded even 28%.

REFERENCES

Bracha Cz. (1998), *Modele reprezentacyjne w badaniu opinii publicznej i marketingu*, EFEICT, Warszawa.

Bracha Cz. (2003), *Kilka uwag o szacowaniu efektu schematu losowania*, [w:] *Metoda reprezentacyjna w badaniach ekonomiczno-społecznych*, Wydawnictwo Akademii Ekonomicznej w Katowicach, Katowice, 13–26.

Goldberger A. S. (1972), *Teoria ekonometrii*, PWE, Warszawa.

Holt D., Smith T. M. E. (1979), *Post-stratification*, J. Roy. Statist. Soc. S. A., **142**, 33–46.

Kish L. (1965), *Survey Sampling*, J. Wiley, New York.

Rao C. R. (1982), *Modele liniowe statystyki matematycznej*, PWN, Warszawa.

Scott A. J., Holt D. (1982), *The effect of two-stage sampling on ordinary least squares methods*, JASA, 848–854.

Theil H. (1979), *Zasady ekonometrii*, PWN, Warszawa.

Wu C. F., Holt D., Holmes D. J. (1988), *The effect of two-stages sampling on the F statistic*, JASA, 150–159.

Czesław Domański

WERYFIKACJA HIPOTEZ DOTYCZĄCYCH PARAMETRÓW MODELU REGRESJI DLA PRÓB NIEPROSTYCH

Problem szacowania parametrów funkcji regresji na podstawie prób nieprostych jest badany z górą od dwudziestu pięciu lat. Przedmiotem badania będzie liniowa funkcja regresji postaci macierzowej: $y = \hat{\beta}x + \varepsilon$, gdzie β jest wektorem nieznanych parametrów, natomiast ε jest składnikiem resztowym.

W przypadku prób nieprostych należy dokonać modyfikacji statystyki testowej, uwzględniając tzw. efekt schematu losowania.

W pracy prezentowane są wyniki badań symulacyjnych, które wskazują na konieczność weryfikacji hipotezy $H_0: \beta = \beta_0$ za pomocą zmodyfikowanego testu *F*.