KAJIAN TENTANG PENGARUH *TWO STAGE CLUSTER SAMPLING* TERHADAP STATISTIK UJI-F

Agung Priyo Utomo (agung@stis.ac.id)
Sekolah Tinggi Ilmu Statistik

ABSTRACT

In regression analysis we make several assumptions about the error term. The following assumptions are often made: 1) the error terms are random variables with mean 0; 2) nonautocorrelation; 3) homoscedasticity, and 4) normality. The assumption of identically and independently distributed (iid) observations that underlies regression procedures is called into question when analyzing complex survey data. Particularly the existence of clusters in two stage samples usually exhibit positive intracluster correlation. If we use Ordinary Least Squares (OLS) procedures to make inferences in regression analysis for two stage cluster samples, we will be faced with a problem. This study aims to know the effect of two stage least squares on the F-Statistic. In general, although OLS procedures are unbiased but not fully efficient for estimation of the regression coefficients. Variance of the OLS estimators for the regression coefficients can be larger than the usual OLS variance expression would indicate. Failure to consider this possibility leads to underestimation of variances, with consequences for confidence intervals and the F-Statistic. The effect of intracluster correlation on the F-Statistic is the distortion of its distribution. The F-Statistic will not follow the Central F distribution anymore. Consequently, the hypothesis testing procedure is invalid.

Keywords: F-Statistic, intracluster correlation, Two Stage Cluster Sampling.

Sensus maupun survei merupakan kegiatan pengumpulan data. Tujuan utama dilakukannya suatu sensus atau survei adalah untuk memperoleh data observasi yang berisi informasi mengenai karakteristik dari populasi (*parameter*) yang akan diteliti. Cochran, W. G. (1977) mendefinisikan populasi adalah kumpulan dari seluruh unit-unit atau elemen-elemen yang termasuk dalam lingkup penelitian. Survei umumnya dilakukan jika banyaknya unit atau elemen populasi yang akan diamati sangat besar, sehingga cukup diambil sebagian dari populasi yang akan diamati.

Untuk pengambilan sebagian unit dari populasi (sampel), terdapat banyak metode yang dapat diterapkan. Untuk survei berskala besar, maka metode yang sesuai untuk digunakan adalah Metode Penarikan Sampel Bertahap Ganda (*Multistage Sampling*), yaitu suatu teknik pengambilan sampel dimana pengambilan sampelnya dilakukan secara bertahap (Cochran, W. G., 1977). Diantara banyak metode yang tergolong dalam *Multistage Sampling*, Metode Penarikan Sampel Dua Tahap (*Two Stage Sampling*) merupakan metode yang paling sederhana. Salah satu metode yang termasuk dalam *Two Stage Sampling* adalah Metode Penarikan Sampel Bergerombol Dua Tahap (*Two Stage Cluster Sampling*). Tahap pertama dalam *Two Stage Sampling* adalah pemilihan *primary sampling unit* (*psu*) dan pada tahap kedua dilakukan pemilihan *secondary sampling unit* (*ssu*). Sebagai contoh, misalkan suatu penelitian dengan unit analisis rumah tangga di Propinsi DKI Jakarta, maka tahap pertama dapat dilakukan pemilihan kecamatan yang ada di wilayah DKI Jakarta sebagai *psu*, dan

selanjutnya pada tahap kedua dilakukan pemilihan rumah tangga pada kecamatan terpilih sebagai *ssu*.

Berbagai metode analisis dapat digunakan untuk memperoleh kesimpulan mengenai data yang dikumpulkan sesuai dengan tujuan yang akan dicapai. Salah satu metode statistik yang sering digunakan dalam menganalisis data adalah analisis regresi. Metode ini digunakan untuk mengetahui bentuk hubungan antar variabel yang dinyatakan dalam suatu model statistik. Umumnya, dalam analisis regresi diberlakukan beberapa asumsi, yaitu kenormalan, homogenitas varian (*homoscedasticity*), dan kebebasan nilai komponen kesalahan (*nonautocorrelation*). Untuk memperoleh hasil terbaik, maka diperlukan suatu pengujian untuk mengetahui apakah variabel-variabel penjelas (*predictor variables*) yang digunakan dapat menjelaskan variasi dalam variabel tak tebas (*response variable*), serta untuk mengetahui apakah asumsi-asumsi yang telah diberlakukan terpenuhi.

Hipotesis mengenai parameter model regresi dapat diuji menggunakan Statistik Uji-t dan Statistik Uji-F. Pengujian akan sahih (*valid*) jika asumsi yang mendasari model terpenuhi. Namun jika data yang digunakan diperoleh dari suatu populasi menggunakan metode *Two Stage Cluster Sampling*, maka biasanya akan memperlihatkan adanya korelasi *intracluster* positif. Keadaan ini pada akhirnya akan mempengaruhi pengujian hipotesis yang dilakukan, terutama pengujian dengan menggunakan Statistik Uji-F.

Berdasarkan permasalahan yang telah diuraikan, maka penelitian ini bertujuan untuk mengetahui bagaimana pengaruh *Two Stage Cluster Sampling* terhadap Statistik Uji-F dalam pengujian terhadap parameter pada suatu model regresi linier.

Sebagaimana yang dikemukakan oleh Cochran, W. G. (1977), *Two Stage Cluster Sampling* merupakan suatu metode penarikan sampel dua tahap dimana pada tahap pertama dilakukan pemilihan sampel gerombol (*cluster*) dari populasi yang terbagi dalam gerombol-gerombol yang disebut sebagai pemilihan *primary sampling unit* (*psu*). Pada tahap kedua, dari *psu* terpilih dilakukan pemilihan elemen-elemen sebagai *secondary sampling unit* (*ssu*).

Untuk menduga nilai parameter dalam suatu persamaan regresi dapat digunakan metode Kuadrat Terkecil Biasa (*Ordinary Least Squares*, OLS). Metode OLS ditemukan pertama kali oleh Carl Friedrich Gauss, seorang ahli matematika Jerman, sehingga sering pula disebut sebagai metode Gauss. Prinsip kerja dari metode OLS adalah meminimalkan jumlah kuadrat komponen kesalahan (*error*).

Dalam notasi matriks, persamaan regresi dinyatakan sebagai berikut:

$$y = X\beta + \varepsilon \qquad (1)$$

dimana:

- y = vektor variabel tak bebas berukuran $n \times 1$, n adalah banyaknya sampel
- X = matrik variabel penjelas berukuran $n \times k$
- β = vektor parameter koefisien regresi yang tidak diketahui, berukuran $k \times 1$
- ε = vektor komponen kesalahan (*error*)

Untuk menduga β, Gauss menerapkan asumsi-asumsi berikut (Draper & Smith, 1981):

1. $E(\varepsilon) = 0$, yaitu nilai harapan dari setiap komponen dalam ε adalah nol.
2. Varian $(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I$, yaitu asumsi adanya kesamaan varian komponen kesalahan (*homoscedasticity*).

3. Matriks $X_{(n\times k)}$ adalah *nonstochastic*, artinya memiliki nilai yang tetap (*fixed*) dari sampel ke sampel
4. Matriks $X_{(n\times k)}$, dimana k < n, mempunyai rank k yang menunjukkan banyaknya vektor kolom yang bebas linier, atau dengan kata lain tidak ada multikolinieritas.
5. Kenormalan distribusi dari komponen kesalahan atau $\varepsilon \sim N(0, \sigma^2 I)$. Asumsi ini digunakan pada pengujian hipotesis dan pembentukan selang kepercayaan (*confidence interval*).

Dengan meminimalkan jumlah kuadrat komponen kesalahan, yaitu meminimalkan $\varepsilon'\varepsilon = (y - X\beta)'(y - X\beta)$), maka akan diperoleh penduga untuk β sebagai berikut:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2}$$

Jika asumsi-asumsi yang diterapkan terpenuhi, maka penduga OLS dari β merupakan penduga linier tak bias terbaik atau sering disebut *best linear unbiased estimator* (BLUE).

Tahap selanjutnya dalam analisis regresi adalah melakukan pengujian hipotesis. Untuk melakukan pengujian tentang parameter koefisien regresi dapat digunakan Statistik Uji-F. Christensen, R. (1984) dan Scott, A. J. dan Holt, D. (1982) merumuskan statistik uji-F sebagai berikut:

$$F(\boldsymbol{\beta}) = \frac{\left\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\right\|^2 \Big/ k}{\left\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right\|^2 \Big/ (n-k)} \tag{3}$$

Statistik uji di atas mengikuti distribusi F dengan derajat bebas k dan n – k, dimana k menyatakan banyaknya parameter di dalam model. Distribusi F adalah suatu distribusi yang merupakan rasio dua variabel acak yang berdistribusi *Chi-Squares* yang saling bebas dibagi dengan masing-masing derajat bebasnya. Sebagaimana dinyatakan oleh Myers, R. H. dan Milton, J. S. (1991), jika U dan V masing-masing merupakan variabel acak berdistribusi *Central Chi-Squares* yang saling bebas dan masing-masing memiliki derajat bebas m dan n, maka variabel acak

$$F_{m,n} = \frac{U \big/ m}{V \big/ n} \tag{4}$$

akan mengikuti distribusi *Central* F dengan derajat bebas m dan n. Apabila salah satu U atau V berditribusi *Noncentral Chi-Squares*, maka variabel acak F di atas akan berdistribusi *Noncentral* F.

METODOLOGI

Untuk mengetahui bagaimana pengaruh *Two Stage Cluster Sampling* terhadap Statistik Uji-F dalam pengujian terhadap parameter pada suatu model regresi linier, penelitian ini menggunakan data simulasi. Data populasi terlebih dahulu dibangkitkan dan dibagi menjadi 5 gerombol. Dari 5 gerombol yang terbentuk akan dipilih 3 gerombol sebagai sampel dengan teknik *simple random sampling without replacement* (SRS WOR). Selanjutnya dari masing-masing gerombol terpilih diambil sampel sebanyak 10 elemen dengan teknik SRS WOR. Data diperlukan untuk membangun model regresi linier sederhana sekaligus melihat pengaruh dari teknik pengambilan sampel yang digunakan terhadap Statistik Uji-F.

HASIL DAN PEMBAHASAN

Korelasi *intracluster* adalah korelasi yang terjadi dalam suatu gerombol (*cluster*) yang biasanya muncul sebagai akibat dari pembentukan gerombol dari unit-unit observasi dalam suatu populasi (Scott, A. J. dan Holt, D., 1982). Korelasi *intracluster* mengukur bagaimana kesamaan unit-unit observasi di dalam suatu gerombol jika dibandingkan dengan populasi yang sangat beragam unit-unitnya. Semakin besar nilai korelasi *intracluster*, maka semakin homogen unit-unit di dalam suatu gerombol. Hal ini bertentangan dengan tujuan pembentukan gerombol sebagaimana dinyatakan oleh Cochran, W. G. (1977), dimana dalam suatu gerombol unit-unitnya diusahakan seheterogen mungkin dan antar gerombol diusahakan sehomogen mungkin, sehingga gerombol yang dibentuk dapat mewakili populasi yang sebenarnya.

Pembentukan gerombol-gerombol dalam populasi akan berdampak pada model regresi pada persamaan (1), yaitu komponen kesalahan merupakan komponen dari pembentukan gerombol dan komponen kesalahan elemen-elemen dalam gerombol itu sendiri. Model regresi untuk elemen ke-i dalam gerombol ke-1 menjadi

$$y_{1i} = \beta_1 + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + \alpha_1 + \varepsilon_{1i} \quad (5)$$

dengan asumsi

$$\alpha_1 \sim N(0,\ \sigma^2_{\alpha_1}) \text{ dan } \varepsilon_{1i} \sim N(0,\ \sigma^2_{\varepsilon_{1i}})$$

di mana:

y_{i1} = nilai variabel tak bebas ke-i pada gerombol ke-1, untuk i = 1, …, n
x_{1j} = nilai variabel bebas ke-j pada gerombol ke-1, untuk j = 1, 2, …, k dimana $x_{11} = 1$
α_1 = komponen kesalahan akibat pembentukan gerombol ke-1
ε_{1i} = komponen kesalahan elemen-elemen dalam gerombol ke-1

Menurut Wu, C. F. J., et al. (1988), korelasi *intracluster* populasi yang dinotasikan dengan ρ dirumuskan sebagai berikut:

$$\rho = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\varepsilon} \quad (6)$$

dimana σ^2_α adalah varian di dalam gerombol dan σ^2_ε merupakan varian antar gerombol.

Estimasi korelasi *intracluster* berdasarkan hasil sampel dapat dilakukan dengan menggunakan rumus berikut:

$$\hat{\rho} = \frac{\hat{\sigma}^2_\alpha}{\hat{\sigma}^2_\alpha + \hat{\sigma}^2_\varepsilon} \quad (7)$$

di mana:

$\hat{\sigma}^2_\varepsilon = \dfrac{1}{c(m-1)} \sum_{i=1}^{c} \left\| \mathbf{e}_l - \bar{\mathbf{e}}_l \mathbf{1} \right\|^2$ = estimasi varian di dalam gerombol

$\hat{\sigma}^2_\alpha = \dfrac{1}{m} \left\{ \dfrac{m}{c-1} \left\| \bar{\mathbf{e}}_l - \bar{\bar{\mathbf{e}}} \mathbf{1} \right\|^2 - \hat{\sigma}^2_\varepsilon \right\}$ = estimasi varian antar gerombol

$\mathbf{e}_l$ = vektor komponen kesalahan pada gerombol ke-*l*

$\bar{\mathbf{e}}_l$ = rata-rata komponen kesalahan pada gerombol ke-*l*

$$\bar{\bar{\mathbf{e}}}\mathbf{1} = \left\{\sum_{l=1}^{c}\bar{\mathbf{e}}_l\right\} \Big/ c$$

1 = vektor satuan

Sedangkan korelasi *intracluster* antar variable bebas dihitung dengan rumus:

$$\rho_x = \frac{1}{m-1}\left[m\frac{m\sum_{l=1}^{c}(\bar{x}_{l.}-\bar{x}_{..})^2}{T_x} - 1 \right] \tag{8}$$

di mana $T_x = \sum_{l=1}^{c} T_{x,l}$ dan $T_{x,l} = \sum_{i=1}^{m_l}(x_{li}-\bar{x}_{l.})^2$.

Pada kasus *Two Stage Cluster Sampling*, yang umumnya akan memunculkan adanya korelasi *intracluster*, maka matriks varian-kovarian komponen kesalahan (ε) pada model regresi menjadi:

$$\text{Var}(\varepsilon) = \sigma^2 \mathbf{V} \tag{9}$$

dimana V adalah bentuk matriks blok diagonal $\mathbf{V} = \bigoplus_{l=1}^{c}\mathbf{V}_l$ dan $\mathbf{V}_l$ merupakan matriks korelasi berukuran $m_l \times m_l$ pada gerombol ke-*l* dengan bentuk:

$$\mathbf{V}_l = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \vdots & 1 \end{bmatrix}_{m_l \times m_l}$$

Untuk melakukan penaksiran parameter pada model regresi, metode OLS dapat diterapkan dengan beberapa asumsi. Sebagaimana dikemukakan oleh Gauss-Markov, pada saat semua asumsi terpenuhi, maka metode OLS akan menghasilkan penaksir parameter yang mempunyai sifat-sifat yang baik, yaitu penaksir tersebut linier, tak bias, dan mempunyai varian yang paling minimum diantara semua kelas penaksir tak bias yang lain (BLUE). Namun menurut Christensen, R. (1984) dan Scott, A. J. dan Holt, D. (1982), bila terjadi korelasi *intracluster* maka penaksir OLS akan memiliki sifat-sifat berikut:

1. Penaksir tersebut tak bias (*unbiased*), yaitu dalam pengambilan sampel yang berulang-ulang (bersyarat pada X yang tetap) nilai rata-ratanya sama dengan nilai populasi.
2. Penaksir tersebut konsisten, yaitu dengan meningkatnya ukuran sampel sampai tak terhingga, penaksir tersebut jatuh ke nilai sebenarnya.
3. Penaksir tersebut kurang efisien baik dalam sampel kecil maupun sampel besar.

Scott, A. J. dan Holt, D. (1982) menyatakan bahwa besarnya nilai efisiensi yang hilang (*loss of efficiency*) dari penaksir OLS dapat dicari dengan membandingkan penaksir OLS dengan penaksir yang diperoleh dengan metode *Generalized Least Squares* (GLS). Formula untuk menghitung hilangnya efisiensi penaksir OLS dapat diturunkan dengan memisalkan e(c), yaitu perbandingan

varian penaksir GLS terhadap penaksir OLS yang masing-masing dikalikan dengan sembarang vektor koefisien c, sehingga diperoleh

$$\frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2} \le e(c) \le 1$$

di mana

$\lambda_1 \ge \lambda_2 \ge \lambda_3 \ge \ldots \ge \lambda_n$ merupakan akar ciri (*eigen value*) dari **V**

$\lambda_n = 1 - \rho$ (untuk $\rho \ge 0$)

$\lambda_1 = 1 + (m_0 - 1)\rho$, dimana m_0 = maximum ($m_1, m_2, \ldots, m_c$)

Dengan demikian dapat disimpulkan bahwa dalam kasus *Two Stage Cluster Sampling*, dimana akan memunculkan adanya korelasi *intracluster*, maka penaksir GLS lebih efisien dibandingkan dengan penaksir OLS. Jika tetap menggunakan metode OLS untuk penaksiran parameter dalam kasus terjadinya korelasi *intracluster*, maka akan membawa beberapa konsekuensi berikut:

1. Jika korelasi *intracluster* dalam penaksir OLS diabaikan, penaksir tersebut tidak efisien jika dibandingkan dengan BLUE. Akibatnya selang kepercayaan (*confidence interval*) yang terbentuk menjadi lebih lebar.
2. $\hat{\sigma}^2$ (penaksir varian komponen kesalahan) akan *underestimate* terhadap σ^2.
3. Penaksir OLS tak bias, namun penaksir tersebut akan memberikan gambaran yang menyimpang dari nilai populasi.
4. Statistik uji-F tidak lagi sahih (*invalid*), dan jika diterapkan akan memberikan kesimpulan yang menyesatkan mengenai keberartian (signifikansi) secara statistik dari koefisien regresi yang ditaksir.

Konsekuensi yang terakhir disebabkan karena Statistik Uji-F yang digunakan dalam pengujian keberartian model regresi sebagaimana dirumuskan pada persamaan (3) diturunkan berdasarkan penaksiran parameter model regresi dengan menggunakan metode OLS, dimana salah satu komponen pembentuk statistik uji tersebut adalah $\hat{\boldsymbol{\beta}}$. Akibat selanjutnya adalah pada distribusi dari Statistik Uji-F. Distribusi dari Statistik Uji-F tidak lagi mengikuti distribusi *Central* F karena pembilang dari statistik uji ini tidak berdistribusi *Central Chi-Squares*, sehingga prosedur pengujian dengan menggunakan statistik uji tersebut menjadi *invalid*.

Wu, C. F. J., et al (1988) menjelaskan bahwa pada kasus terjadinya korelasi *intracluster*, Statistik Uji-F tidak lagi memiliki tingkat signifikansi yang sebenarnya, yaitu sebesar α, dan *confidence ellipsoid* yang terbentuk akan menyimpang dari cakupan yang seharusnya. Dijelaskan lebih lanjut bahwa jika dimisalkan $\boldsymbol{\delta} = \sigma^{-1}\mathbf{V}^{-\frac{1}{2}}\boldsymbol{\varepsilon}$, maka Statistik Uji-F dapat ditulis sebagai berikut:

$$F = \frac{\boldsymbol{\delta}'\mathbf{V}^{\frac{1}{2}}\mathbf{P}\mathbf{V}^{\frac{1}{2}}\boldsymbol{\delta}/k}{\boldsymbol{\delta}'\mathbf{V}^{\frac{1}{2}}(\mathbf{I}-\mathbf{P})\mathbf{V}^{\frac{1}{2}}\boldsymbol{\delta}/(n-k)} \qquad (10)$$

dimana $\boldsymbol{\delta}' = (\delta_1, \ldots, \delta_n)$ mengikuti distribusi normal bebas (*independent*) dengan rata-rata 0 dan varian 1. Tingkat signifikansi yang sebenarnya dari Statistik Uji-F adalah

$$\Pr\{F \ge F_{\alpha;k,n-k}\} = \Pr\{\boldsymbol{\delta}'\mathbf{V}^{\frac{1}{2}}[\mathbf{P} - k(n-k)^{-1}F_{\alpha;k,n-k}(\mathbf{I}-\mathbf{P})]\mathbf{V}^{\frac{1}{2}}\boldsymbol{\delta} \ge 0\}$$

dan cakupan dari *confidence ellipsoid* yang sebenarnya adalah

$$\Pr\{F \le F_{\alpha;k,n-k}\} = \Pr\{\boldsymbol{\delta}'\mathbf{V}^{\frac{1}{2}}[\mathbf{P} - k(n-k)^{-1}F_{\alpha;k,n-k}(\mathbf{I}-\mathbf{P})]\mathbf{V}^{\frac{1}{2}}\boldsymbol{\delta} \le 0\}$$

Suatu cara yang dapat digunakan untuk melihat penyimpangan dari Statistik Uji-F adalah dengan melihat nilai tr(PV)/k, suatu konstanta dari χ^2_k. Jika nilai tr(PV)/k tidak sama dengan 1, berarti distribusi dari Statistik Uji-F tidak mengikuti distribusi *Central* F (Wu, C. F. J., et al., 1988). Untuk jumlah sampel (n) yang besar, dengan tingkat signifikansi sebesar α_1 maka Statistik Uji-F akan memiliki tingkat signifikansi sebenarnya minimal sebesar $\alpha 2$, dimana $\alpha_2 > \alpha_1$, jika

$$tr(\mathbf{PV})/k \le \frac{F_{\alpha_1(k,n-k)}}{F_{\alpha_2(k,n-k)}} \qquad (11)$$

Bukti Empiris

Untuk memperoleh gambaran yang lebih jelas mengenai permasalahan yang telah diuraikan, maka digunakan model regresi linier sederhana

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

dimana

Y_i merupakan nilai variabel tak bebas untuk sampel ke-i
X_i merupakan nilai variabel bebas untuk sampel ke-i
β_0 dan β_1 merupakan parameter model regresi
ε_i merupakan komponen error (kesalahan) ke-i yang bersifat stokastik
i = 1, 2, …, n

Tahap awal pada two stage cluster sampling adalah pemilihan *psu*, yaitu pemilihan sampel sebanyak 3 gerombol dari populasi gerombol dengan teknik *simple random sampling without replacement* (SRS WOR). Tahap berikutnya, dari masing-masing gerombol terpilih diambil sampel sebanyak 10 elemen dengan teknik SRS WOR. Penggunaan sampel dengan jumlah yang sama untuk masing-masing gerombol dimaksudkan agar memperjelas pengaruh korelasi *intracluster* terhadap metode OLS dan Statistik Uji-F tanpa dipengaruhi oleh perbedaan ukuran sampel untuk masing-masing gerombol. Untuk melihat pengaruh dari korelasi *intracluster* dengan besar yang berbeda, maka pengambilan sampel dilakukan sebanyak 10 kali.

Pengaruh korelasi *intracluster* dapat diketahui melalui tahapan berikut:

1. Melakukan pendugaan terhadap model regresi dengan metode OLS pada kasus *Two Stage Cluster Sampling*.
2. Menghitung penduga korelasi *intracluster* menggunakan formula (7) dan (8).
3. Menghitung tr(PV).
4. Mendapatkan nilai F_{α_2} menggunakan formula (11) dan membandingkannya dengan nilai F_{tabel} pada α_1 = 5%, dimana $F_{0.05;1,29} = 4{,}183$. Dengan demikian akan diketahui tingkat signifikansi yang sebenarnya (α_2).

Berdasarkan langkah-langkah di atas, maka diperoleh beberapa informasi yang terangkum dalam Tabel 1.

Tabel 1. Tingkat signifikansi sebenarnya (α_2) dari 10 kali pengambilan sampel pada saat tingkat signifikansi yang digunakan (α_1) = 5%, c = 3, dan m = 10

Sampel	$\hat{\rho}$	$\hat{\rho}_x$	tr(PV)	$F_{\alpha_2,1,29}$	α_2 (%)
I	0,2250	0,2478	1,5320	2,7300	10,93
II	0,0212	0,4060	1,0520	3,9780	5,56
III	0,0555	0,2408	1,0800	3,8720	5,87
IV	0,0549	0,0568	1,0190	4,1050	5,20
V	0,0504	0,4059	1,1840	3,5300	7,04
VI	0,0533	0,0257	1,0080	4,1498	5,09
VII	0,1830	0,1029	1,1140	3,7570	6,24
VIII	0,0294	0,2350	1,0410	4,0170	5,45
IX	0,0700	0,0330	1,0140	4,1260	5,15
X	0,0796	0,1699	1,0810	3,8690	5,88

Tabel di atas memperlihatkan bahwa semakin besar korelasi intracluster antar sisaan ($\hat{\rho}$) maupun variabel bebas ($\hat{\rho}_x$), maka semakin besar pula penyimpangan tingkat signifikansi yang terjadi. Pada saat tingkat signifikansi yang digunakan sebesar 5%, pada kasus sampel yang diambil dengan *two stage cluster sampling*, tingkat signifikansi atau tingkat kesalahan yang sebenarnya terjadi selalu lebih besar. Hal ini harus diwaspadai karena akan berdampak pada kesahihan pengujian yang dilakukan. Nilai tr(PV) dapat dijadikan indikator yang baik untuk mengetahui tingkat penyimpangan Statistik Uji-F sebagai akibat dari adanya korelasi *intracluster*. Semakin jauh nilai tr(PV) dari angka 1, maka makin tinggi penyimpangan yang terjadi.

Untuk memperjelas pengaruh korelasi *intracluster* terhadap Statistik Uji-F serta untuk mengetahui korelasi *intracluster* mana yang paling dominan, maka dilakukan simulasi dengan beberapa tingkatan korelasi yang mungkin terjadi. Hasilnya tercantum dalam tabel 2.

Tabel 2. Tingkat signifikansi sebenarnya (%) dari Statistik Uji-F dengan tingkat signifikansi (α_1) 5% pada beberapa tingkatan korelasi *intracluster* (c = 3 dan m = 10)

$\hat{\rho}_x$	$\hat{\rho}$						
	0,00	0,01	0,05	0,10	0,20	0,25	0,30
0,00	5,00 (1,00)	5,24 (1,05)	7,79 (1,23)	11,24 (1,45)	18,31 (1,45)	21,71 (2,13)	24,94 (2,35)
0,01	5,00 (1,00)	5,25 (1,05)	7,82 (1,23)	11,31 (1,46)	18,44 (1,45)	21,98 (2,14)	25,14 (2,36)
0,05	5,00 (1,00)	5,27 (1,05)	7,95 (1,24)	11,59 (1,47)	18,99 (1,45)	22,52 (2,18)	25,88 (2,42)
0,10	5,00 (1,00)	5,30 (1,05)	8,12 (1,25)	11,94 (1,50)	19,67 (1,45)	23,35 (2,24)	26,82 (2,49)
0,20	5,00 (1,00)	5,36 (1,05)	8,46 (1,27)	12,66 (1,54)	21,04 (1,45)	24,94 (2,35)	28,61 (2,62)
0,30	5,00 (1,00)	5,42 (1,06)	8,80 (1,29)	14,08 (1,59)	21,04 (1,45)	26,50 (2,46)	30,34 (2,76)
0,40	5,00 (1,00)	5,48 (1,06)	9,14 (1,32)	14,20 (1,63)	23,66 (1,45)	28,02 (2,58)	32,00 (2,89)

*) Nilai dalam tanda kurung merupakan nilai tr(PV)/k

Beberapa informasi penting dari tabel di atas adalah:

a. Pada saat $\hat{\rho}$ = 0, tidak ada perubahan pada tingkat signifikansi dari Statistik Uji-F meskipun $\hat{\rho}_x \neq 0$.
b. Pada saat $\hat{\rho} \neq 0$ namun $\hat{\rho}_x = 0$, ada pengaruh dari korelasi intracluster tersebut terhadap Statistik Uji-F. Semakin besar $\hat{\rho}$ semakin besar pula tingkat signifikansi yang sebenarnya terjadi, meskipun $\hat{\rho}_x = 0$.
c. Pada $\hat{\rho}_x$ yang sama namun $\hat{\rho} \neq 0$ dan berubah-ubah, terlihat penyimpangan yang lebih besar dibandingkan penyimpangan yang terjadi pada kondisi $\hat{\rho}$ yang sama namun $\hat{\rho}_x \neq 0$ dan berubah-ubah.

Berdasarkan beberapa informasi di atas, maka dapat disimpulkan bahwa pengaruh yang paling dominan terhadap penyimpangan tingkat signifikansi yang sebenarnya dari Statistik Uji-F berasal dari korelasi *intracluster* sisaan.

KESIMPULAN DAN SARAN

Penerapan *Two Stage Cluster Sampling* umumnya akan memunculkan adanya korelasi *intracluster* antar sisaan dan antar variabel bebas. Korelasi *intracluster* tersebut akan berdampak pada prosedur inferensia yang dilakukan dengan metode OLS. Statistik Uji-F yang digunakan dalam pengujian akan menyimpang dari distribusi F sebagai akibat dari varian sisaan OLS yang *underestimate* pada kasus terjadinya korelasi *intracluster*. Hal ini akan membawa konsekuensi bahwa tingkat signifikansi yang sebenarnya dari Statistik Uji-F akan lebih besar dari tingkat signifikansi yang telah ditetapkan sebelumnya. Penyimpangan tingkat signifikansi yang sebenarnya dari Statistik Uji-F tersebut sebagian besar disebabkan oleh korelasi *intracluster* sisaan. Pada akhirnya prosedur pengujian dengan menggunakan statistik tersebut menjadi tidak sahih.

Dengan demikian, apabila data yang akan dianalisis berasal dari *Two Stage Cluster Sampling* maka perlu diperiksa terlebih dahulu besar kecilnya korelasi *intracluster* yang terjadi, terutama korelasi *intracluster* antar sisaan.

REFERENSI

Christensen, R. (1984). A note on ordinary least squares methods for two stage sampling. *Journal of The American Statistical Association,* 79, p. 720 – 721.

Cochran, W.G. (1977). *Sampling techniques.* 3rd edition. New York: John Wiley and Sons. (Terjemahan).

Draper, N.R. & Smith, H. (1981). *Applied regression analysis*. 2nd edition. New York: John Wiley. (Terjemahan).

Myers, R.H. & Milton, J.S. *A first course in the theory of linear statistical models*. Boston: PWS-KENT Publishing Company.

Scott, A.J. & Holt, D. (1982). The effect of two stage sampling on ordinary least squares methods. *Journal of The American Statistical Association,* 77, p. 848 – 854.

Wu, C.F.J., Holt, D., & Holmes, D.J. (1988). The effect of two stage sampling on the F statistic. *Journal of The American Statistical Association,* 83, p. 150 – 159.