ДЗ-02 РЕШЕНИЕ

Задача 1.
Дан набор наблюдений $(x1_i,\ x2_i,\ x3_i,\ y_i),\ i=1....,n$. Рассматривается регрессионная модель

$$y_i=\beta_1 x1_i+\beta_2 x2_i+\beta_3 x3_i+\varepsilon_i,\quad V(\varepsilon_i)=\sigma^2,\ E(\varepsilon_i)=0,\quad \text{все } \varepsilon_i \text{ независимы.}$$

Поскольку похоже, что имеется мультиколлинеарность, рассматривается следующий метод оценивания:

$$f(\beta)=\sum_{i=1}^{n}(y_i-\beta_1 x1_i-\beta_2 x2_i-\beta_3 x3_i)^2+\lambda\cdot\sum_{j=1}^{3}\beta_j^2 \xrightarrow[\beta]{} \min$$

Результат $\hat\beta(\lambda)$ зависит от заданного параметра λ. При $\lambda=0$ полученная оценка совпадает с оценкой МНК: $\hat\beta(0)=\hat\beta_{LS}=(X'X)^{-1}X'y$.

(a) Найдите выражение для $\hat\beta(\lambda)$ в матричном виде.

(b) Является ли оценка $\hat\beta(\lambda)$ несмещенной?

(c) Найдите выражение для матрицы ковариаций $V(\hat\beta(\lambda))$ в матричном виде.

(d) Существует ли значение параметра λ, такое, что оценка компоненты вектора параметров $\hat\beta_j(\lambda)$ имеет дисперсию меньше, чем у МНК-оценки $\hat\beta_{j,LS}$?

(e) Пусть данные приведены в Таблице 1

Таблица 1

x1	x2	x3	y
1	0.67	0.32	3.030211
1	0.98	0.03	3.317133
1	0.31	0.68	3.050206
1	0.21	0.80	2.882887
1	0.70	0.29	3.083099
1	0.01	0.99	2.845725
1	0.49	0.50	2.765274
1	0.30	0.69	2.583751
1	0.50	0.51	2.814222
1	0.34	0.65	2.999865

Эти данные сгенерированы по (истинной) модели $y_i=2\cdot x1_i+x2_i+x3_i+\varepsilon_i,\ \ \sigma^2=0.09$.

(e1) Найдите МНК оценки коэффициентов β_j, $j=1,2,3$, и точные дисперсии этих оценок. Значимы ли эти коэффициенты?

(e2) Постройте на одном рисунке графики зависимости $\hat\beta_j(\lambda)$, $j=1,2,3$ от λ в диапазоне $\lambda\in(0.0001,\,100)$. Используйте логарифмические шкалы.

(e3) Постройте на одном рисунке графики зависимости $MSE\left(\hat\beta_j(\lambda)\right)$ (рассчитанные с учетом истинного значения $\sigma^2=0.009$), $j=1,2,3$ от λ в диапазоне $\lambda\in(0.0001,\,100)$. Используйте логарифмические шкалы.

(e4) Прокомментируйте полученные результаты.

Решение

(a) $f(\beta)=(y-X\beta)'(y-X\beta)'+\lambda\beta'\beta$ условия первого порядка

$$\frac{\partial f}{\partial\beta}=-2X'y+2X'X\beta+2\lambda\beta=0$$, откуда $\hat\beta(\lambda)=(X'X+\lambda I)^{-1}X'y$.

(b) $E\left(\hat\beta(\lambda)\right)=(X'X+\lambda I)^{-1}X'Ey=(X'X+\lambda I)^{-1}X'X\beta=\beta-\lambda(X'X+\lambda I)^{-1}\beta$, оценка смещенная.

(c) $V\left(\hat\beta(\lambda)\right)=V\left((X'X+\lambda I)^{-1}X'y\right)=(X'X+\lambda I)^{-1}X'\cdot\sigma^2 I\cdot X(X'X+\lambda I)^{-1}=$

$$=\sigma^2(X'X+\lambda I)^{-1}X'X(X'X+\lambda I)^{-1}=\sigma^2\left((X'X+\lambda I)^{-1}-\lambda(X'X+\lambda I)^{-2}\right)$$

(d) при $\lambda \geq 0$ имеем $V(\hat{\beta}_{LS})^{-1} = \sigma^{-2} X'X \leq \sigma^{-2}(X'X + \lambda I) = V(\hat{\beta}(\lambda))^{-1}$, т.е. $V(\hat{\beta}(\lambda)) \leq V(\hat{\beta}_{LS})$.

(e)

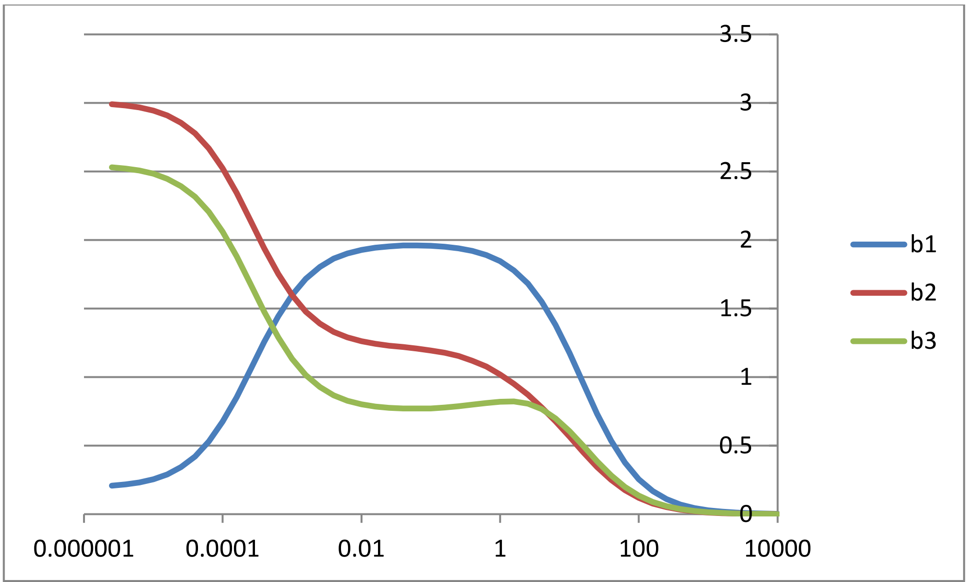
МНК оценка:

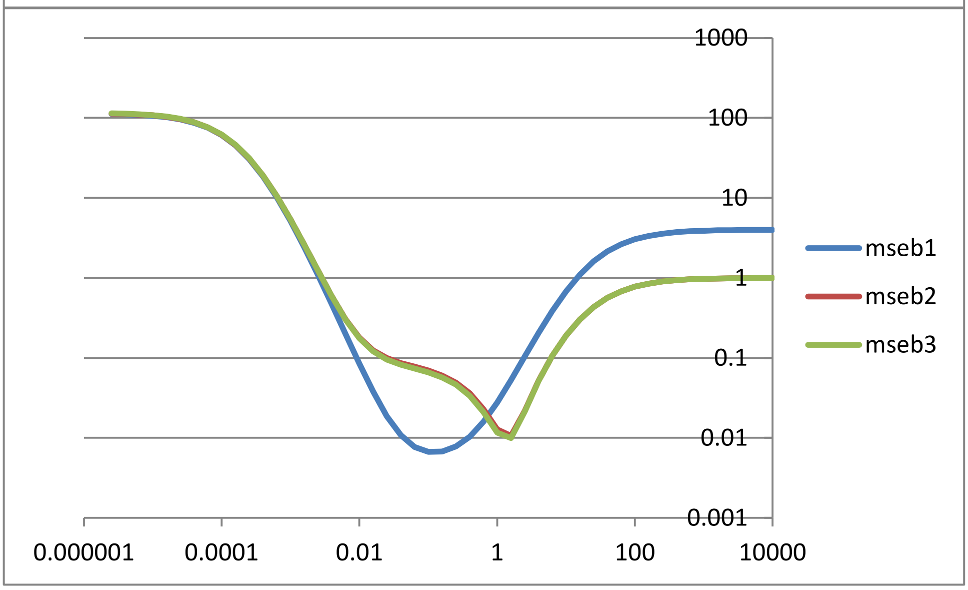
```
      Source |       SS           df       MS      Number of obs   =        10
-------------+----------------------------------   F(3, 7)         =    949.11
       Model |  86.4330475         3  28.8110158   Prob > F        =    0.0000
    Residual |  .212491012         7  .030355859   R-squared       =    0.9975
-------------+----------------------------------   Adj R-squared   =    0.9965
       Total |  86.6455385        10  8.66455385   Root MSE        =    .17423

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          x1 |   .1908276   6.124555     0.03   0.976    -14.29144     14.6731
          x2 |   3.006852   6.134351     0.49   0.639    -11.49858    17.51229
          x3 |   2.546373   6.151259     0.41   0.691    -11.99904    17.09179
------------------------------------------------------------------------------
```

Точные значения стандартных ошибок равны

```
10.546
10.563
10.592
```





Выокое значение R2 и большие стандартные ошибки коэффициентов.
$Corr(x_2, x_3) = -0.9994$. Мультиколлинеарность.

При $\lambda \approx 1.58$ MSE оценок коэффициентов $\beta_2,\ \beta_3$ наименьшие при этом оценки $\beta_1,\ \beta_2,\ \beta_3$ равны соответственно $1.78,\ 0.95,\ \ 0.82$.

Задача 2

Пусть есть модель с пространственной зависимостью: $y_i = \mu + \varepsilon_i$. $E\varepsilon_i = 0,\ \ V(\varepsilon_i) = \sigma^2,\ \ \varepsilon_i \sim iid(0,\sigma^2)$, $i = 1,\ldots,n$. Ошибки коррелированы: $Corr(\varepsilon_i,\varepsilon_k) = \rho \ge 0,\ \ i \ne k$.

(a) Найдите МНК-оценку $\hat{\mu}$ параметра μ. Верно ли что $\lim\limits_{n\to\infty} V(\hat{\mu}) = 0$ при $\rho > 0$? Сравните дисперсии оценок при $n = 10,\ \rho = 0$ и при $n = 10,\ \rho = 0.05$. Какое количество наблюдений m надо взять при $\rho = 0.05$ чтобы достичь той же точности оценки, что и при $n = 10,\ \rho = 0$. Интерпретируйте результат как потерю информации при наличии зависимости.

(b) Возьмите оценку s^2 дисперсии ошибки $V(\varepsilon_i) = \sigma^2$. Вычислите $E(s^2)$ при $\rho = 0$ и при $\rho \ne 0$.

(c) Пусть ρ известно. Найдите GLS оценку $\hat{\mu}_{GLS}$ параметра μ.

Решение

(a) $X = \iota' = (1 \ldots 1)'$, $y = (y_1 \ldots y_n)'$. Имеем $\hat{\mu} = (X'X)^{-1}X'y = \bar{y}$.

$$V(\hat{\mu}) = V(\bar{y}) = V\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2}V\left(\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n} V(y_i) + \sum_{i,j=1,i\ne j}^{n} Cov(y_i,y_j)\right) = \frac{1}{n^2}\left(n\sigma^2 + n(n-1)\rho\sigma^2\right) =$$

$$= \frac{\sigma^2}{n^2}\left(n + n(n-1)\rho\right) = \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2 \ge \frac{\sigma^2}{n}.\ \ \lim_{n\to\infty} V(\hat{\mu}) = \rho\sigma^2 > 0.$$

При $n = 10,\ \rho = 0$, $V(\hat{\mu}) = \dfrac{\sigma^2}{10}$; при $n = 10,\ \rho = 0.05$, $V(\hat{\mu}) = 1.45\dfrac{\sigma^2}{10}$.

Должно быть $\dfrac{\sigma^2}{m}\left(1 + 0.05(m-1)\right) \le \dfrac{\sigma^2}{10}$, или $0.05 + \dfrac{0.95}{m} \le 0.01$, или $m \ge \dfrac{0.95}{0.05} = 19$.

Поскольку ε_i и ε_j коррелированы, то при известном ε_i уже известна часть информации о ε_j.

(b) $s^2 = \dfrac{e'e}{n-k} = \dfrac{e'e}{n-1} = \dfrac{1}{n-1}\sum_{1}^{n}(y_i - \bar{y})^2$.

$$E\left(\sum_{1}^{n}(y_i - \bar{y})^2\right) = E\left(\sum_{1}^{n}(y_i^2 - 2\bar{y}y_i + \bar{y}^2)\right) = nE(y_i^2) + nE(\bar{y}^2) - 2nE(\bar{y}y_1) =$$

$$= n(\mu^2 + \sigma^2) + n\left(\mu^2 + \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2\right) - 2n\frac{1}{n}\sum_{i=1}^{n} E(y_i y_1) =$$

$$= n(\mu^2 + \sigma^2) + n\left(\mu^2 + \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2\right) - 2\left(Ey_1^2 + \sum_{i=2}^{n} E(y_i y_1)\right) =$$

$$= n(\mu^2 + \sigma^2) + n\left(\mu^2 + \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2\right) - 2\left(\mu^2 + \sigma^2 + (n-1)(\mu^2 + \rho\sigma^2)\right) =$$

$$= \sigma^2\left(n + 1 + (n-1)\rho - 2 - 2(n-1)\rho\right) = \sigma^2(n-1)(1-\rho).$$ Поэтому $E(s^2) = (1-\rho)\sigma^2$.

(c) $V(\varepsilon) = \sigma^2\left((1-\rho)I_n + \rho\iota\iota'\right) = \sigma^2(1-\rho)\left(I_n + \gamma\iota\iota'\right) = \sigma^2(1-\rho)\Omega,\ \ \gamma = \dfrac{\rho}{1-\rho}$.

Найдем обратную матрицу в виде $\Omega^{-1} = I_n + \delta\iota\iota'$.

$I_n = (I_n + \gamma\iota\iota')(I_n + \delta\iota\iota') = I_n + (\gamma + \delta + \delta\gamma)\iota\iota'$, откуда $\delta = -\dfrac{\gamma}{1+\gamma} = -\rho$.

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y = \left(\iota'(I_n - \rho\iota\iota')\iota\right)^{-1}\iota'(I_n - \rho\iota\iota')y = (n - \rho n^2)^{-1}(n\bar{y} - \rho n^2\bar{y}) = \bar{y} = \hat{\beta}.$$

Задача 3.

Предположим, данные порождены моделью $y_t = \alpha + \beta x_t + \varepsilon_t,\ \ t = 1,\ldots,n$, удовлетворяеющей условиям классической регрессии. $\hat{\alpha},\ \hat{\beta}$ — оценки метода наименьших квадратов. Оценка $\tilde{\beta}$ получена по методу наименьших квадратов при дополнительном (вообще говоря, неверном) предположении, что $\alpha = 0$.

(а) Найдите МНК-оценку $\tilde{\beta}$. При каких условиях она является несмещенной оценкой параметра β?

(б) Найдите дисперсию оценки $\tilde{\beta}$, сравните ее с дисперсией оценки $\hat{\beta}$.

(в) Обсудите, какую из двух оценок лучше использовать.

Решение

(а) МНК-оценка параметра β в регрессии y на x при ограничении $\alpha = 0$ получается минимизацией (по β) суммы $\sum_{t=1}^{n}(y_t - \beta x_t)^2$. Дифференцируя по β и приравнивая производную нулю, получаем $\tilde{\beta} = \frac{\sum_{t=1}^{n} x_t y_t}{\sum_{t=1}^{n} x_t^2}$.

Учитывая, что $y_t = \alpha + \beta x_t + \varepsilon_t$, получаем:
$$\tilde{\beta} = \beta + \alpha\frac{\sum_{t=1}^{n} x_t}{\sum_{t=1}^{n} x_t^2} + \frac{\sum_{t=1}^{n} x_t\varepsilon_t}{\sum_{t=1}^{n} x_t^2} \qquad (1)$$

Из (1) следует, что $\mathrm{E}(\tilde{\beta}) = \beta + \alpha\frac{\sum_{t=1}^{n} x_t}{\sum_{t=1}^{n} x_t^2}$, поэтому оценка $\tilde{\beta}$ будет несмещенной, если либо α = 0 (что очевидно), либо $\sum_{t=1}^{n} x_t = 0$ (ортогональность вектора *x* вектору констант).

(б) Из (1) в силу некоррелированности ошибок ε сразу следует, что $\mathrm{V}(\tilde{\beta}) = \frac{\sigma^2}{\sum_{t=1}^{n} x_t^2}$. Как известно, $\mathrm{V}(\hat{\beta}) = \frac{\sigma^2}{\sum_{t=1}^{n} x_t^2 - n\bar{x}^2}$, откуда вытекает неравенство $\mathrm{V}(\tilde{\beta}) < \mathrm{V}(\hat{\beta})$. Это не противоречит теореме Гаусса-Маркова, поскольку оценка $\tilde{\beta}$, вообще говоря, смещена.

(в) Можно попытаться сравнить эти оценки по среднеквадратичной ошибке. Легко видеть, что $\mathrm{MSE}(\hat{\beta}) = \mathrm{V}(\hat{\beta}) = \frac{\sigma^2}{\sum_{t=1}^{n} x_t^2 - n\bar{x}^2}$, $\mathrm{MSE}(\tilde{\beta}) = \frac{\sigma^2}{\sum_{t=1}^{n} x_t^2} + \alpha^2\left(\frac{\sum_{t=1}^{n} x_t}{\sum_{t=1}^{n} x_t^2}\right)^2 = \frac{\sigma^2}{\sum_{t=1}^{n} x_t^2} + \alpha^2\frac{n^2\bar{x}^2}{\left(\sum_{t=1}^{n} x_t^2\right)^2}$ Точный и полный анализ ситуаций, когда одна из ошибок больше другой, достаточно труден, но грубые прикидки показывают, что при малом разбросе переменной *x* дисперсия $\mathrm{V}(\hat{\beta})$ становится очень большой, и тогда целесообразно использовать оценку $\tilde{\beta}$.

Задача 4

Процесс, порождающий данные (DGP) имеет вид

$y_i = \beta_0 x_i + u_i,\ u_i = x_i\varepsilon_i;\ x_i \sim i.i.d.N(0,1);\quad \varepsilon_i \sim i.i.d.N(0,1).$

Кроме того, все x_i и ε_j независимы для всех i, j.

Замечание. Если $\xi \sim N(0,\sigma^2)$, то $E(\xi) = E(\xi^3) = 0,\ E(\xi^2) = \sigma^2,\ E(\xi^4) = 3\sigma^4$.

(a) Покажите, что ошибка u_i условно гетероскедастична (найдите $V(u\,|\,x)$).

(b) Найдите $p\lim\frac{1}{N}X'X$.

(c) Найдите $\sigma_0^2 = V(u_i)$. (Учитывая все случайные переменные в модели).

(d) Найдите $p\lim\frac{1}{N}X'\Omega_0 X = \lim\frac{1}{N}E\left(X'\Omega_0 X\right)$, где $\Omega_0 = diag\{V(u_i\,|\,x_i)\}$.

(e) Используя полученные результаты найдите асимптотическую дисперсию МНК оценки $\hat{\beta}_{OLS}$, (т.е. дисперсию предельного распределения $\sqrt{N}(\hat{\beta}_{OLS} - \beta_0)$) по стандартной формуле, игнорируя гетероскедастичность. Ответ должен быть численным.

(f) Теперь найдите дисперсию предельного распределения $\sqrt{N}(\hat{\beta}_{OLS} - \beta_0)$ учитывая гетероскедастичность. Ответ должен быть численным.

(g) Как полученное различие в результатах (e) и (f) соответствует вашим ожиданиям?

(h) Приведите код STATA, который генерирует данные в соответствии с DGP в условии, с $N = 100$.
Оценивая регрессию с предположением гомоскедастичности найдите оценку для (e).
Оценивая регрессию с робастными стандартными ошибками найдите оценку для (f).
Сравните с вашими теоретическими результатами.
Повторите для $N = 1000$.

Решение

(a) The error u is conditionally heteroskedastic, since $V(u\,|\,x) = V(x\varepsilon\,|\,x) = x^2V(\varepsilon\,|\,x) = x^2V(\varepsilon) = x^2\cdot 1 = x^2$. which depends on the regressor x.

(b) For scalar regressor $N^{-1}X'X = N^{-1}\sum_i x_i^2$. Here x_i^2 are i.i.d. with mean 1 (since $E(x_i^2) = E(x_i^2) - E(x_i)^2 = V(x_i) = 1$). Applying a LLN $p\lim N^{-1}X'X = p\lim N^{-1}\sum_i x_i^2 = E(x_i^2) = 1$, so $M_{xx} = 1$.

(c) $V(u) = V(x\varepsilon) = E[(x\varepsilon)^2] - [E(x\varepsilon)]^2 = E(x^2)E(\varepsilon^2) - [E(x)E(\varepsilon)]^2 = V(x)V(\varepsilon) - 0 = 1$,
where use independence of x and ε and fact that here $Ex = E\varepsilon = 0$.

(d) For scalar regressor and diagonal Ω_0, $N^{-1}X'\Omega_0 X = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2 x_i^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^4$. Here x_i^4 are i.i.d. with mean 3 (fourth central moment of normal is $3\sigma^4 = 3 \cdot 1 = 3$). Applying a LLN,

$p\lim N^{-1}X'\Omega_0 X = p\lim \frac{1}{N}\sum_{i=1}^{N} x_i^4 = E(x_i^4) = 3,$ so $M_{x\Omega_0 x} = 3$.

(e) Default OLS result $\sqrt{N}(\hat{\beta}_{OLS} - \beta) \xrightarrow{D} N(0, \sigma^2 M_{xx}^{-1}) = N(0,1)$.

(f) White OLS result $\sqrt{N}(\hat{\beta}_{OLS} - \beta) \xrightarrow{D} N(0, \sigma^2 M_{xx}^{-1} M_{x\Omega x} M_{xx}^{-1}) = N(0,3)$.

(g) Yes. Expect that failure to control for conditional heteroskedasticity when should control for it will lead to inconsistent standard errors, though a priori the direction of the inconsistency is not known. That is the case here. What is unusual compared to many applications is that there is a big di¤erence in this example - the true variance is three times the default estimate and the true standard errrors are $\sqrt{3}$ times larger.