

## 4. Regresja wielokrotna liniowa

W rozdziale 2 zajmowaliśmy się taką sytuacją, w której w populacji generalnej  $\pi$  rozważaliśmy tylko dwie zmienne: zmienną losową  $Y$  i zmienną losową lub rzeczywistą  $X$ . Obecny rozdział poświęcimy takiej sytuacji, w której w populacji generalnej  $\pi$  obserwować będziemy zmienną losową  $Y$  i  $k$  zmiennych losowych lub rzeczywistych  $X_i$  ( $i = 1, 2, \dots, k$ ).

O zmiennej losowej  $Y$  założymy, że jest to zmienna losowa normalna:

$$Y \sim N\left(m(x_1, x_2, \dots, x_k); \sigma_{y/x_1, x_2, \dots, x_k}\right) . \quad (4.1)$$

O wartości oczekiwanej zmiennej losowej  $Y$  założymy dalej, że jest funkcją liniową zmiennych  $X_i$  postaci:

$$m(x_1, x_2, \dots, x_k) = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k = b_0 + \sum_{i=1}^k b_i x_i . \quad (4.2)$$

Wariancja  $\sigma_{y/x_1, x_2, \dots, x_k}^2$  odchyłeń od modelu funkcji regresji jest stała dla dowolnej kombinacji wartości zmiennych losowych  $X_i$ .

Parametry modelu (4.2) nie są znane i muszą być estymowane na podstawie odpowiedniej próby losowej. Oznaczmy elementy tej próby losowej jako  $(y_j, x_{1j}, \dots, x_{kj})$ , gdzie  $j = 1, 2, \dots, n$  jest wskaźnikiem powtórzeń (replikacji). Zgodnie z modelem (4.2) dowolną obserwację empiryczną możemy przedstawić jako:

$$y_j = b_0 + b_1 x_{1j} + b_2 x_{2j} + \dots + b_k x_{kj} + e_j \quad \text{dla } j = 1, 2, \dots, n . \quad (4.3)$$

Symbol  $e_j$  oznacza resztę, różnicę między wartością obserwowaną  $y_j$  a wartością teoretyczną  $\hat{y}_j$  wynikającą z modelu:

$$e_j = y_j - (b_0 + b_1 x_{1j} + b_2 x_{2j} + \dots + b_k x_{kj}) = y_j - \hat{y}_j . \quad (4.4)$$

Kryterium estymacji nieznanymi parametrów modelu możemy sformułować tak: chcemy tak dobrać parametry modelu, aby różnice między wartościami obserwowanymi a teoretycznymi były jak najmniejsze. W sensie matematycznym warunek ten sprowadza się do zminimalizowania funkcji  $s$ :

$$s = \sum_{j=1}^n e_j^2 = \sum_{j=1}^n \left[ y_j - (b_0 + b_1 x_{1j} + b_2 x_{2j} + \dots + b_k x_{kj}) \right]^2 = \text{minimum} \quad (4.5)$$

Wyznaczenie minimum funkcji  $s$  określonej wzorem (4.5) wymaga rozwiązania układu  $k+1$  równań normalnych. Problem można trochę uprościć przekształcając równość (4.3) w taki sposób, aby wyeliminować stałą regresji  $b_0$ :

$$y_j - \bar{y} = b_1(x_{1j} - \bar{x}_1) + b_2(x_{2j} - \bar{x}_2) + \dots + b_k(x_{kj} - \bar{x}_k) + e_j \quad (4.6)$$

gdzie

$$b_0 = \bar{y} - (b_1\bar{x}_1 + b_2\bar{x}_2 + \dots + b_k\bar{x}_k) = \bar{y} - \sum_{i=1}^k b_i\bar{x}_i. \quad (4.7)$$

Uwzględniając wzór (4.6) kryterium estymacji można zapisać następująco:

$$\begin{aligned} s &= \sum_{j=1}^n e_j^2 = \\ &= \sum_{j=1}^n \left[ y_j - \bar{y} - (b_1(x_{1j} - \bar{x}_1) + b_2(x_{2j} - \bar{x}_2) + \dots + b_k(x_{kj} - \bar{x}_k)) \right]^2 = \text{minimum} \end{aligned} \quad (4.8)$$

Znalezienie minimum funkcji  $s$  określonej wzorem (4.8) wymaga wyznaczenia  $k$  pochodnych cząstkowych funkcji  $s$  względem parametrów  $b_i$  (gdzie  $i = 1, 2, \dots, k$ ), a następnie przyrównania ich do zera i rozwiązanie powstałego w ten sposób układu równań:

$$\frac{\partial s}{\partial b_i} = -2 \sum_{j=1}^n \left[ y_j - (b_1(x_{1j} - \bar{x}_1) + b_2(x_{2j} - \bar{x}_2) + \dots + b_k(x_{kj} - \bar{x}_k)) \right] (x_{ij} - \bar{x}_i) = 0 \quad (4.9)$$

Otrzymany w wyniku zastosowania wzoru (4.9) układ równań normalnych możemy zapisać w tradycyjnej postaci:

$$\begin{cases} \hat{b}_1 \text{var } x_1 + \hat{b}_2 \text{cov } x_1 x_2 + \dots + \hat{b}_k \text{cov } x_1 x_k &= \text{cov } x_1 y \\ \hat{b}_1 \text{cov } x_2 x_1 + \hat{b}_2 \text{var } x_2 + \dots + \hat{b}_k \text{cov } x_2 x_k &= \text{cov } x_2 y \\ \vdots & \vdots \\ \hat{b}_1 \text{cov } x_k x_1 + \hat{b}_2 \text{cov } x_k x_2 + \dots + \hat{b}_k \text{var } x_k &= \text{cov } x_k y \end{cases} \quad (4.10)$$

W układzie równań określonym wzorem (4.10) w miejsce parametrów modelu w populacji generalnej  $b_i$  wprowadzono już ich oceny z próby  $\hat{b}_i$ .

Układ równań normalnych przedstawiony wzorem (4.10) wygodniej jest zapisać w notacji macierzowej:

$$\mathbf{V}\hat{\mathbf{B}} = \mathbf{C} \quad (4.11)$$

gdzie

$$\mathbf{V}_{(k \times k)} = \begin{bmatrix} \text{var } x_1 & \text{cov } x_1 x_2 & \dots & \text{cov } x_1 x_k \\ \text{cov } x_2 x_1 & \text{var } x_2 & \dots & \text{cov } x_2 x_k \\ \vdots & \vdots & \dots & \vdots \\ \text{cov } x_k x_1 & \text{cov } x_k x_2 & \dots & \text{var } x_k \end{bmatrix} \quad (4.12)$$

$$\hat{\mathbf{B}}_{(k \times 1)} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \vdots \\ \hat{b}_k \end{bmatrix} \quad \mathbf{C}_{(k \times 1)} = \begin{bmatrix} \text{COV } x_1 y \\ \text{COV } x_2 y \\ \vdots \\ \text{COV } x_k y \end{bmatrix} \quad (4.13)$$

Macierz  $\mathbf{V}$ , macierz współczynników przy niewiadomych, jest macierzą kwadratową stopnia  $k$ , jest to macierz symetryczna zawierająca na głównej przekątnej sumy kwadratów odchyłeń zmiennych niezależnych  $X_i$ , a poza główną przekątną sumy iloczynów odchyłeń tych zmiennych. Jeżeli macierz  $\mathbf{V}$  jest macierzą nieosobliwą (czyli jej wyznacznik jest różny od zera), to istnieje macierz odwrotna do macierzy  $\mathbf{V}$  oznaczana symbolem  $\mathbf{V}^{-1}$ . Mnożąc równanie (4.11) lewostronnie przez macierz odwrotną do  $\mathbf{V}$  otrzymujemy:

$$\mathbf{V}\hat{\mathbf{B}} = \mathbf{C} \mid \cdot \mathbf{V}^{-1} \Rightarrow \mathbf{V}^{-1}\mathbf{V}\hat{\mathbf{B}} = \mathbf{V}^{-1}\mathbf{C} \Rightarrow \mathbf{I}\hat{\mathbf{B}} = \mathbf{V}^{-1}\mathbf{C} \Rightarrow \hat{\mathbf{B}} = \mathbf{V}^{-1}\mathbf{C} \quad (4.14)$$

Po wyestymowaniu parametrów  $\hat{b}_i$  (dla  $i = 1, 2, \dots, k$ ) z równania (4.14) wyznaczamy ocenę parametru  $b_0$  z wzoru (4.7):

$$\hat{b}_0 = \bar{y} - \sum_{i=1}^k \hat{b}_i \bar{x}_i. \quad (4.15)$$

Ostatecznie ocena funkcji regresji z próby ma postać:

$$\hat{m}(x_1, x_2, \dots, x_k) = \hat{b}_0 + \sum_{i=1}^k \hat{b}_i x_i. \quad (4.16)$$

Parametr  $b_0$  będziemy nazywać, podobnie jak w regresji liniowej, stałą regresji. Parametry  $b_1, b_2, \dots, b_k$  będziemy nazywać **częstkowymi** współczynnikami regresji.

## 4.1. Badanie istotności modelu

Po wyestymowaniu funkcji regresji z próby musimy sobie odpowiedzieć na pytanie, czy nasze założenie o tym, że wartość oczekiwana zmiennej losowej  $Y$  jest funkcją liniową zmiennych  $X_i$  jest prawdziwe.

Nie będzie takiej zależności wtedy, gdy wszystkie cząstkowe współczynniki regresji będą jednocześnie równe zero. Tym samym hipotezie o **istotności regresji wielokrotnej liniowej** możemy nadać postać:

$$H_0 : \bigwedge_i b_i = 0 \quad (\text{dla } i = 1, 2, \dots, k) \quad (4.17)$$

Tabela analizy wariancji dla weryfikacji  $H_0 : \bigwedge_i b_i = 0$  wobec  $H_1 : \bigvee_i b_i \neq 0$ 

| Zmiennosc | Stopnie swobody   | Suma kwadratow odchylen                           | Średni kwadrat odchylen   | $F$ empiryczne            |
|-----------|-------------------|---|---------------------------|---------------------------|
| Modelu    | $v_R = k$         | $SS_R = \sum_{i=1}^k \hat{b}_i \text{cov } x_i y$ | $MS_R = \frac{SS_R}{v_R}$ | $F_R = \frac{MS_R}{MS_E}$ |
| Resztowa  | $v_E = n - k - 1$ | $SS_E = SS_T - SS_R$                              | $MS_E = \frac{SS_E}{v_e}$ |                           |
| Całkowita | $v_T = n - 1$     | $SS_T = \text{var } y$                            |                           |                           |

Hipotezę  $H_0 : \bigwedge_i b_i = 0$  będziemy odrzucać na korzyść  $H_1 : \bigvee_i b_i \neq 0$  wtedy, gdy wartość empiryczna statystyki  $F$  Fishera będzie większa od odpowiedniej wartości krytycznej  $F_{emp.} > F_{\alpha, v_R, v_E}$  lub krytyczny poziom istotności ( $p$ -value) będzie mniejszy od przyjętego poziomu istotności alfa. Merytorycznie sformułujemy wniosek, że **istnieje istotna liniowa zależność** między zmienną losową  $Y$  a zmiennymi niezależnymi  $X_i$  (co najmniej jedną z nich).

Jeżeli wartość empiryczna statystyki  $F$  Fishera jest niewiększa od odpowiedniej wartości krytycznej lub  $p$ -value jest większe od przyjętego alfa, to nie mamy podstaw do odrzucenia  $H_0$ . Merytorycznie oznacza to, że **nie istnieje liniowy związek** między zmienną losową  $Y$  a zmiennymi  $X_i$ . W tej sytuacji wartość oczekiwana zmiennej losowej  $Y$  jest stała i równa wartości średniej.

Wróćmy jednak do sytuacji, gdy hipotezę  $H_0$  odrzucimy. Proszę zwrócić uwagę, że odrzucenie hipotezy zerowej daje stosunkowo mało informacji. Jedynie co wiemy, to to, że **co najmniej jeden cząstkowy współczynnik regresji jest różny od zera**. Podobnie jak w przypadku szczegółowych porównań w analizie wariancji musimy przeprowadzić dalsze szczegółowe badania zmierzające do ustalenia, **które** cząstkowe współczynniki regresji są różne od zera.

Teoretycznie sprawa jest stosunkowo prosta: wystarczy zweryfikować serię  $k$  hipotez zerowych o istotności cząstkowych współczynników regresji postaci:

$$H_{0i} : b_i = 0 \text{ wobec } H_{1i} : b_i \neq 0 \text{ dla } i = 1, 2, \dots, k \quad . \quad (4.18)$$

Hipotezy te weryfikujemy testem  $t$ -Studenta, gdzie funkcja testowa określona jest wzorem:

$$t_i = \frac{\hat{b}_i}{S_{\hat{b}_i}} = \frac{\hat{b}_i}{\sqrt{S_{y/x_1, x_2, \dots, x_k}^2 \cdot v^{ii}}} \quad (4.19)$$

Błąd standardowy oceny  $i$ -tego, cząstkowego współczynnika regresji, wyznaczamy jako pierwiastek kwadratowy iloczynu średniego kwadratu odchyleń od modelu regresji pomnożonego przez element diagonalny macierzy odwrotnej do macierzy  $\mathbf{V}$ :

$$S_{\hat{b}_i} = \sqrt{S_{y/x_1, x_2, \dots, x_k}^2 \cdot v^{ii}} = \sqrt{MS_E \cdot v^{ii}} \quad (4.20)$$

Statystyka określona wzorem (4.19) ma, przy prawdziwości hipotezy zerowej, rozkład  $t$ -Studenta z liczbą stopni swobody  $v_E = n - k - 1$ . W sytuacji, gdy wartość empiryczna statystyki  $t$  znajduje się w obszarze krytycznym dla  $H_0$ , to hipotezę odrzucamy. Tym samym  $i$ -ta zmienna niezależna **powinna** pozostać w modelu funkcji regresji.

W sytuacji odwrotnej (wartość empiryczna statystyki  $t$  znajduje się w obszarze dopuszczalnym dla  $H_0$ ) nie mamy podstaw do jej odrzucenia, co powinno być równoważne z tym, że  $i$ -ta zmienna niezależna  $x_i$  **może** być usunięta z modelu funkcji regresji.

Jak zasygnalizowałem na poprzedniej stronie teoretycznie problem określenia, które zmienne niezależne powinny pozostać w modelu funkcji regresji powinien być prosty. W praktyce jest jednak zupełnie inaczej, a cała trudność wynika z faktu, że oceny z próby poszczególnych cząstkowych współczynników regresji **nie są niezależne**. Tym samym również statystyki  $t$  zdefiniowane wzorem (4.19) nie są niezależne, co w konsekwencji może prowadzić do zupełnie (pozornie) niezrozumiałych rozstrzygnięć.

Może tak się zdarzyć, że testem  $F$  Fishera odrzucimy hipotezę o istotności regresji, czyli co najmniej jedna zmienna niezależna powinna pozostać w modelu funkcji regresji. Weryfikując jednak serię hipotez o istotności kolejnych cząstkowych współczynników regresji możemy nie być w stanie odrzucić żadnej z nich, co powinno sugerować, że wszystkie zmienne powinny być usunięte z modelu funkcji regresji. Może tak się dzieć wtedy, gdy zmienne niezależne są silnie wzajemnie z sobą powiązane (co można stwierdzić wyznaczając choćby współczynniki korelacji liniowych między parami zmiennych). W takiej sytuacji decydowanie o tym, które zmienne mają pozostać w modelu w oparciu o weryfikację serii hipotez określonych wzorem (4.18) nie wchodzi w rachubę.

Działanie takie, jak przedstawione powyżej jest poprawne wtedy tylko, gdy zmienne niezależne (objasniające)  $X$  są **wzajemnie niezależne**, czyli wtedy, gdy macierz  $V$  jest macierzą **diagonalną**. W każdej innej sytuacji (a tak jest najczęściej) musimy szukać innej metody pozwalającej na optymalne dobranie zmiennych w modelu. Przed jej wprowadzeniem przedstawimy jeszcze miary dobroci dopasowania modelu.

## 4.2. Dokładność dopasowania modelu

Podobnie jak w przypadku regresji liniowej jednej zmiennej niezależnej możemy wprowadzić współczynnik determinacji  $R^2$  określający stopień dopasowania modelu funkcji regresji do empirycznych wartości zmiennej losowej  $Y$ :

$$R^2 = \frac{\sum_{i=1}^k \hat{b}_i \text{cov } x_i y}{\text{var } y} . \quad (4.21)$$

Współczynnik determinacji  $R^2$  informuje nas o tym, jaka część zmienności całkowitej zmiennej losowej  $Y$  jest wyjaśniona przez zmienne niezależne uwzględnione w modelu funkcji regresji.

Współczynnik determinacji  $R^2$  przyjmuje swoje wartości z przedziału  $\langle 0; 1 \rangle$ , z tym, że najczęściej wyrażamy go w procentach  $\langle 0\%; 100\% \rangle$ .

Kolejną miarą dobroci dopasowania modelu jest współczynnik korelacji wielokrotnej  $R$  definiowany jako pierwiastek kwadratowy ze współczynnika determinacji:

$$R = \sqrt{R^2} = \sqrt{\frac{\sum_{i=1}^k \hat{b}_i \text{cov } x_i y}{\text{var } y}} \quad (4.22)$$

Współczynnik korelacji wielokrotnej  $R$  przyjmuje swoje wartości z przedziału  $\langle 0; 1 \rangle$ , im model jest lepiej dopasowany, tym  $R$  jest bliższe wartości 1.

Istotnym parametrem określającym dobroć dopasowania modelu jest średni kwadrat odchyłeń wartości obserwowanych i teoretycznych (reszt)  $MS_E = S_{y/x_1, x_2, \dots, x_k}^2$ . Im ten średni kwadrat odchyłeń jest mniejszy, tym model jest lepiej dopasowany. Wielkość  $MS_E$  wpływa także na błędy estymacji parametrów modelu oraz błąd wartości regresyjnej i błąd predykcji.

### 4.3. Dokładność oceny parametrów modelu

Parametry modelu (4.2) są szacowane z próby losowej, tym samym ich oceny obarczone są pewnym błędem. Jak wiemy (zobacz wzór (4.20)) ocenę błędu standardowego cząstkowego współczynnika regresji znajdziemy z wzoru:

$$S_{\hat{b}_i} = \sqrt{MS_E \cdot v^{ii}} = \sqrt{S_{y/x_1, x_2, \dots, x_k}^2 \cdot v^{ii}} \quad \text{dla } i = 1, 2, \dots, k \quad (4.23)$$

co pozwala w konsekwencji na zbudowanie przedziału ufności dla prawdziwej wartości tego współczynnika w populacji generalnej:

$$b_i \in < \hat{b}_i - t_{\alpha, n-k-1} S_{\hat{b}_i}; \hat{b}_i + t_{\alpha, n-k-1} S_{\hat{b}_i} > \quad z \quad P = 1 - \alpha \quad . \quad (4.24)$$

Wyznaczenie oceny błędu standardowego stałej regresji jest trochę bardziej skomplikowane:

$$S_{\hat{b}_0} = \sqrt{MS_E \cdot A_0} = \sqrt{S_{y/x_1, x_2, \dots, x_k}^2 \cdot A_0} \quad (4.25)$$

gdzie wielkość  $A_0$  wyznaczana jest z wzoru:

$$A_0 = \frac{1}{n} \left[ 1 + \frac{1}{n} \mathbf{D}_1^T \mathbf{V}^{-1} \mathbf{D}_1 \right] \quad . \quad (4.26)$$

Wektor  $\mathbf{D}_1^T$  występujący we wzorze (4.26) jest wektorem sum zmiennych niezależnych wyznaczonych na podstawie  $n$ -elementowej próby losowej:

$$\mathbf{D}_1^T = \left[ \sum_{j=1}^n x_{1j} \quad \sum_{j=1}^n x_{2j} \quad \dots \quad \sum_{j=1}^n x_{kj} \right] \quad . \quad (4.27)$$

Przedział ufności dla stałej regresji w populacji generalnej znajdziemy z wzoru:

$$b_0 \in < \hat{b}_0 - t_{\alpha, n-k-1} S_{\hat{b}_0}; \hat{b}_0 + t_{\alpha, n-k-1} S_{\hat{b}_0} > \quad z \quad P = 1 - \alpha \quad . \quad (4.28)$$

Interpretacja zbudowanych zgodnie ze wzorami (4.24) i (4.28) przedziałów ufności dla cząstkowych współczynników regresji i stałej regresji jest standardowa: zbudowany przedział liczbowy pokrywa nieznaną wartość parametru z prawdopodobieństwem  $1 - \alpha$ .

Przykład 3. Na podstawie poniższych wyników oszacujemy liniową funkcję regresji opisującą zależność indywidualnej wydajności pracy ( $y$ ) od stażu pracy ( $x_1$ ) i wieku ( $x_2$ ) pracowników.

| $j$     | $y_j$ | $x_{1j}$ | $x_{2j}$ |
|---------|-------|----------|----------|
| 1       | 84    | 1        | 19       |
| 2       | 44    | 0,5      | 21       |
| 3       | 94    | 2,5      | 25       |
| 4       | 113   | 9        | 35       |
| 5       | 64    | 3,5      | 25       |
| 6       | 110   | 6        | 47       |
| 7       | 125   | 7,5      | 48       |
| 8       | 94    | 7        | 28       |
| 9       | 126   | 7        | 45       |
| 10      | 121   | 11       | 40       |
| 11      | 76    | 2        | 23       |
| 12      | 58    | 2        | 23       |
| 13      | 51    | 3        | 21       |
| 14      | 118   | 9,5      | 31       |
| 15      | 83    | 4        | 25       |
| 16      | 125   | 5        | 49       |
| 17      | 112   | 8,5      | 46       |
| 18      | 91    | 8        | 27       |
| 19      | 116   | 11       | 43       |
| 20      | 122   | 12       | 35       |
| Sumy    | 1927  | 120      | 656      |
| Średnie | 96,35 | 6        | 32,8     |

Problem sprowadza się do wyestymowania metodą najmniejszych kwadratów parametrów modelu:

$$m(x_1, x_2) = b_0 + b_1x_1 + b_2x_2 .$$

Układ równań normalnych będzie miał postać:

$$\begin{bmatrix} \text{var } x_1 & \text{cov } x_1x_2 \\ \text{cov } x_1x_2 & \text{var } x_2 \end{bmatrix} \cdot \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} \text{cov } x_1y \\ \text{cov } x_2y \end{bmatrix}$$

stąd jednym z pierwszych naszych zadań będzie wyznaczenie potrzebnych elementów macierzy  $\mathbf{V}$  i wektora  $\mathbf{C}$  oraz  $\bar{x}_1$ ,  $\bar{x}_2$ ,  $\bar{y}$  i  $\text{var } y$ .

Niech  $\mathbf{Z}$  oznacza macierz odchyłeń zmiennych empirycznych od ich średnich. Iloczyn macierzy transponowanej do  $\mathbf{Z}$  przez macierz  $\mathbf{Z}$  tworzy macierz  $\mathbf{V}_0$  o strukturze blokowej, gdzie każdy z bloków zawiera interesująca nas liczbę, wektor czy macierz:

$$\mathbf{Z}^T \mathbf{Z} = \mathbf{V}_0 = \begin{bmatrix} \text{var } y & \mathbf{C}^T \\ \mathbf{C} & \mathbf{V} \end{bmatrix} \quad (4.29)$$

Wszelkie niezbędne obliczenia wykonamy w Excelu. Poniżej pokazałem arkusz z wpisanymi od komórki A1 danymi empirycznymi z naszego przykładu. Od komórki H1 wpisana jest macierz transponowana odchyłeń od średnich ( $\mathbf{Z}^T$ ).

|    | A     | B   | C    | D      | E       | F       | G | H                     | I        | J      | K        |
|----|-------|-----|------|--------|---------|---------|---|-----------------------|----------|--------|----------|
| 1  | y     | x1  | x2   | y-ysr  | x1-xsr1 | x2-xsr2 |   | y-ysr                 | -12,35   | -52,35 | -2,35    |
| 2  | 84    | 1   | 19   | -12,35 | -5      | -13,8   |   | x1-xsr1               | -5       | -5,5   | -3,5     |
| 3  | 44    | 0,5 | 21   | -52,35 | -5,5    | -11,8   |   | x2-xsr2               | -13,8    | -11,8  | -7,8     |
| 4  | 94    | 2,5 | 25   | -2,35  | -3,5    | -7,8    |   |                       |          |        |          |
| 5  | 113   | 9   | 35   | 16,65  | 3       | 2,2     |   |                       |          |        |          |
| 6  | 64    | 3,5 | 25   | -32,35 | -2,5    | -7,8    |   | Macierz V0            |          |        |          |
| 7  | 110   | 6   | 47   | 13,65  | 0       | 14,2    |   | 13448,6               | 1407,5   | 4409,4 |          |
| 8  | 125   | 7,5 | 48   | 28,65  | 1,5     | 15,2    |   | 1407,5                | 241,5    | 446,0  |          |
| 9  | 94    | 7   | 28   | -2,35  | 1       | -4,8    |   | 4409,4                | 446,0    | 2087,2 |          |
| 10 | 126   | 7   | 45   | 29,65  | 1       | 12,2    |   |                       |          |        |          |
| 11 | 121   | 11  | 40   | 24,65  | 5       | 7,2     |   | Macierz V             |          |        | Wektor C |
| 12 | 76    | 2   | 23   | -20,35 | -4      | -9,8    |   | 241,5                 | 446,0    |        | 1407,5   |
| 13 | 58    | 2   | 23   | -38,35 | -4      | -9,8    |   | 446,0                 | 2087,2   |        | 4409,4   |
| 14 | 51    | 3   | 21   | -45,35 | -3      | -11,8   |   |                       |          |        |          |
| 15 | 118   | 9,5 | 31   | 21,65  | 3,5     | -1,8    |   | Macierz odwrotna do V |          |        |          |
| 16 | 83    | 4   | 25   | -13,35 | -2      | -7,8    |   | 0,00684               | -0,00146 |        |          |
| 17 | 125   | 5   | 49   | 28,65  | -1      | 16,2    |   | -0,00146              | 0,000791 |        |          |
| 18 | 112   | 8,5 | 46   | 15,65  | 2,5     | 13,2    |   |                       |          |        |          |
| 19 | 91    | 8   | 27   | -5,35  | 2       | -5,8    |   |                       |          |        |          |
| 20 | 116   | 11  | 43   | 19,65  | 5       | 10,2    |   |                       |          |        |          |
| 21 | 122   | 12  | 35   | 25,65  | 6       | 2,2     |   |                       |          |        |          |
| 22 | 1927  | 120 | 656  |        |         |         |   |                       |          |        |          |
| 23 | 96,35 | 6   | 32,8 |        |         |         |   |                       |          |        |          |

Rysunek 7. Fragment arkusza kalkulacyjnego PodstawyEkonometriiPrz2.xls<sup>4</sup> z częściowymi wynikami obliczeń.

Mając macierz  $\mathbf{V}^{-1}$  oraz wektor  $\mathbf{C}$  znajdujemy oszacowania cząstkowych współczynników regresji:

$$\hat{\mathbf{B}} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} 0,00684 & -0,00146 \\ -0,00146 & 0,000791 \end{bmatrix} \cdot \begin{bmatrix} 1407,5 \\ 4409,4 \end{bmatrix} = \begin{bmatrix} 3,1826 \\ 1,4325 \end{bmatrix}.$$

<sup>4</sup> do pobrania z [wszim-sochaczew.edu.pl/download/ekonometria](http://wszim-sochaczew.edu.pl/download/ekonometria)

Możemy już oszacować stałą regresji:

$$\hat{b}_0 = 96,35 - (3,1826 \cdot 6 + 1,4325 \cdot 32,8) = 30,2677.$$

Mamy wyestymowany model funkcji regresji z próby, musimy teraz sprawdzić, czy jest to model istotny, co sprowadza się do zweryfikowania hipotezy zerowej o tym, że wszystkie cząstkowe współczynniki regresji są równe zero. Jak wiemy hipotezę tę będziemy weryfikować testem  $F$  Fishera, jednak przed zestawieniem tabeli analizy wariancji musimy jeszcze doliczyć brakujące elementy.

$$SS_R = 3,1826 \cdot 1407,5 + 1,4325 \cdot 4409,4 = 10796,1$$

$$SS_E = 13448,6 - 10796,1 = 2652,5$$

Tabela analizy wariancji dla weryfikacji  $H_0 : \bigwedge_i b_i = 0$  wobec  $H_1 : \bigvee_i b_i \neq 0$

| Zmienność | Stopnie swobody | Suma kwadratów odchyłeń | Średni kwadrat odchyłeń | $F$ empiryczne | $p$ -value |
|-----------|-----------------|-------------------------|-------------------------|----------------|------------|
| Modelu    | 2               | 10796,1                 | 5398,00                 | 34,596         | 1,017E-06  |
| Resztowa  | 17              | 2652,5                  | 156,03                  |                |            |
| Całkowita | 19              | 13448,6                 |                         |                |            |

Z powyższej tabeli wynika, że  $p$ -value jest mniejsze od przyjętego domyślnie poziomu istotności ( $\alpha = 0,05$ ), tym samym hipotezę zerową odrzucamy. Merytorycznie oznacza to, że istnieje istotny liniowy związek między zmienną losową  $Y$  a co najmniej jedną ze zmiennych niezależnych  $X$ .

Przed zbadaniem istotności cząstkowych współczynników regresji wyznaczymy jeszcze miary dobroci dopasowania modelu, czyli współczynnik determinacji i korelacji wielokrotnej:

$$R^2 = \frac{10796,1}{13448,6} = 0,803 = 80,3\% \quad R = \sqrt{0,803} = 0,896$$

Współczynnik determinacji możemy zinterpretować następująco: zmienność indywidualnej wydajności pracy jest w 80,3% wyjaśniona zmiennymi niezależnymi uwzględnionymi w modelu funkcji regresji. Oznacza to jednocześnie, że prawie 20% tej zmienności nie zostało wyjaśnione estymowanym modelem regresji (co może oznaczać, że model jest źle skonstruowany zarówno co do postaci jak i uwzględnionych zmiennych – zagadnieniom tego typu będą poświęcone dalsze rozdziały tego skryptu).

Przejdziemy teraz do zbadania istotności cząstkowych współczynników regresji, przy czym zaczniemy ten etap pracy od wyznaczenia błędów standardowych ocen obu współczynników. Wykorzystując  $MS_E = 156,03$  z tabeli analizy wariancji oraz elementy diagonalne macierzy odwrotnej do  $\mathbf{V}$  mamy:

$$S_{\hat{b}_1} = \sqrt{156,03 \cdot 0,00684} = 1,0331$$

$$S_{\hat{b}_2} = \sqrt{156,03 \cdot 0,000791} = 0,3514$$

Możemy już wyznaczyć wartości empiryczne statystyki  $t$ -Studenta dla weryfikacji hipotez  $H_0 : b_i = 0$  wobec  $H_1 : b_i \neq 0$  (dla  $i = 1, 2$ ):

$$t_1 = \frac{3,1826}{1,0331} = 3,081$$

$$t_2 = \frac{1,4325}{0,3514} = 4,076$$

Przyjmując, że  $\alpha = 0,05$  odczytujemy z tablic statystycznych (w Excelu mamy funkcję Rozkład.T.Odw) wartość krytyczną tej statystyki przy 17 stopniach swobody:  $t_{\alpha=0,05;v=17} = 2,11$ . Jak widzimy, obie wartości empiryczne statystyki  $t$ -Studenta znajdują się w obszarze krytycznym dla hipotez  $H_0$ , tym samym odrzucamy te hipotezy na korzyść alternatyw.

Podobnie jak w przypadku testu  $F$  w analizie wariancji weryfikację obu hipotez zerowych możemy przeprowadzić także w oparciu o krytyczny poziom istotności (warunek: musimy dysponować co najmniej Excelem). Na liście funkcji statystycznych tego arkusza znajdziemy funkcję Rozkład.T, jej zastosowanie daje następujące wartości krytycznego poziomu istotności ( $p$ -value):

dla  $t_1 = 3,081$  mamy  $p$ -value = 0,0068 ,

dla  $t_2 = 4,076$  mamy  $p$ -value = 7,9E-04.

Jak widzimy dla obu hipotez zerowych krytyczny poziom istotności jest mniejszy od przyjętego  $\alpha$ , co upoważnia nas do odrzucenia hipotez zerowych na korzyść  $H_1$ .

Merytorycznie będziemy wnioskować następująco: istnieje istotna liniowa zależność między indywidualną wydajnością pracy a stażem pracy ( $x_1$ ) i wiekiem pracowników ( $x_2$ ) oszacowana modelem:

$$\hat{m}(x_1, x_2) = 30,2677 + 3,1826x_1 + 1,4325x_2 .$$

Ocenom cząstkowych współczynników regresji możemy nadać następującą interpretację:

- 1) zwiększenie stażu pracy pracownika o jedną jednostkę spowoduje, przy ustalonym wieku pracownika, średnie zwiększenie indywidualnej wydajności pracy o 3,1826 jednostki,
- 2) zwiększenie wieku pracownika o jedną jednostkę spowoduje, przy ustalonym stażu pracy, średnie zwiększenie indywidualnej wydajności pracy o 1,4325 jednostki.

Zakończymy przykład 3 wyznaczeniem przedziałów ufności dla wszystkich trzech parametrów modelu (stałej regresji i obu cząstkowych współczynników regresji). Mamy już wyznaczone błędy standardowe ocen cząstkowych współczynników regresji, pozostało nam wyznaczenie błędu dla stałej regresji. Korzystając z wzoru (4.25) kolejno mamy:

$$\frac{1}{20} \mathbf{D}_1^T = \frac{1}{20} [120 \quad 656] = [6,0 \quad 32,8]$$

$$\frac{1}{20} \mathbf{D}_1^T \mathbf{V}^{-1} = [6,0 \quad 32,8] \cdot \begin{bmatrix} 0,00684 & -0,00146 \\ -0,00146 & 0,000791 \end{bmatrix} = [-0,00690 \quad 0,01719]$$

$$\left( \frac{1}{20} \mathbf{D}_1^T \mathbf{V}^{-1} \right) \mathbf{D}_1 = [-0,00690 \quad 0,01719] \cdot \begin{bmatrix} 120 \\ 656 \end{bmatrix} = 10,448$$

$$A_0 = \frac{1}{20} (1 + 10,448) = 0,5724$$

$$S_{\hat{b}_0} = \sqrt{250,18 \cdot 0,5724} = 9,4505$$

Pozostałe obliczenia zestawimy w formie tabelarycznej, przy czym przedziały ufności wyznaczymy przy 5% poziomie istotności.

| Oszacowanie parametru | Błąd standardowy | Dolna granica | Górna granica |
|-----------------------|------------------|---------------|---------------|
| $\hat{b}_0$           | 30,2677          | 10,3288       | 50,2065       |
| $\hat{b}_1$           | 3,1826           | 1,0030        | 5,3622        |
| $\hat{b}_2$           | 1,4325           | 0,6911        | 2,1739        |

Wyestymowane przedziały ufności można zinterpretować następująco:

- a) z 95% zaufaniem możemy oczekiwać, że stała regresji w populacji będzie nie mniejsza niż 10,3288, lecz nie większa niż 50,2065.
- b) z 95% zaufaniem mamy prawo oczekiwać, że cząstkowy współczynnik regresji stojący przy zmiennej opisującej staż pracy będzie nie mniejszy niż 1,0030, ale

nie większy niż 5,3622 (proszę zauważyć, że przedział ten nie zawiera zera, co jest zgodne z weryfikacją hipotezy o istotności tego współczynnika).

- c) z 95% zaufaniem mamy prawo oczekiwać, że cząstkowy współczynnik regresji stojący przy zmiennej opisującej wiek pracownika będzie nie mniejszy niż 0,6911, ale nie większy niż 2,1739 (proszę zauważyć, że przedział ten nie zawiera zera, co jest zgodne z weryfikacją hipotezy o istotności tego współczynnika).

#### 4.4. Predykcja (prognoza) w regresji wielokrotnej liniowej

Podobnie jak w przypadku regresji liniowej jednej zmiennej niezależnej wyestymowany model funkcji regresji można wykorzystać do wyznaczenia teoretycznej wartości zmiennej losowej  $Y$  dla ustalonego wektora wartości zmiennych niezależnych  $X_i$ .

Zgodnie z przyjętym modelem średnią wartość zmiennej losowej  $Y$  dla ustalonych wartości zmiennych niezależnych  $\mathbf{x}_0 = [x_{10} \ x_{20} \ \dots \ x_{k0}]$  znajdziemy z wzoru:

$$\hat{m}(\mathbf{x}_0) = [1 \ \mathbf{x}_0] \cdot \begin{bmatrix} \hat{b}_0 \\ \hat{\mathbf{B}} \end{bmatrix} = \hat{b}_0 + \sum_{i=1}^k \hat{b}_i x_{i0} . \quad (4.30)$$

Wyznaczona zgodnie z powyższym wzorem wartość regresyjna jest oczywiście losowa, bo **losowe są oceny parametrów modelu**. Standardowy błąd estymacji wartości regresyjnej możemy wyznaczyć z wzoru:

$$S_{\hat{m}(\mathbf{x}_0)} = \sqrt{MS_E [1 \ \mathbf{x}_0] \mathbf{V}_0^{-1} [1 \ \mathbf{x}_0]^T} \quad (4.31)$$

gdzie macierz  $\mathbf{V}_0^{-1}$  jest macierzą blokową postaci:

$$\mathbf{V}_0^{-1} = \begin{bmatrix} A_0 & \mathbf{D}_2 \\ \mathbf{D}_2^T & \mathbf{V}^{-1} \end{bmatrix} . \quad (4.32)$$

Wyrażenie (liczba)  $A_0$  zostało już wcześniej określone wzorem (4.26), a wektor  $\mathbf{D}_2$  znajdziemy z wzoru:

$$\mathbf{D}_2 = -\frac{1}{n} \mathbf{D}_1 \mathbf{V}^{-1} \quad (4.33)$$

gdzie wektor  $\mathbf{D}_1$  był już określony wzorem (4.27) jako wektor sum obserwacji dla zmiennych niezależnych.

Wykorzystując wartość regresyjną określoną wzorem (4.30) oraz błąd standardowy tej oceny dany wzorem (4.31) budujemy przedział ufności dla wartości regresyjnej:

$$m(\mathbf{x}_0) \in \langle \hat{m}(\mathbf{x}_0) - t_{\alpha, n-k-1} S_{\hat{m}(\mathbf{x}_0)}; \hat{m}(\mathbf{x}_0) + t_{\alpha, n-k-1} S_{\hat{m}(\mathbf{x}_0)} \rangle \text{ z } P = 1 - \alpha . \quad (4.34)$$

Przejdziemy teraz do prognozowania nie wartości średniej zmiennej losowej  $Y$ , lecz do prognozowania pojedynczej realizacji tej zmiennej, a to jest właśnie przedmiotem klasycznej predykcji. Zgodnie z modelem liniowym wartość tę wyznaczmy z wzoru:

$$y_{\mathbf{x}_0} = [1 \quad \mathbf{x}_0] \cdot \begin{bmatrix} b_0 \\ \mathbf{B} \end{bmatrix} + e . \quad (4.35)$$

a jej najlepszym estymatorem jest wartość regresyjna  $\hat{m}(\mathbf{x}_0)$ .

Błąd prognozy pojedynczej realizacji zmiennej losowej  $Y$  (błąd predykcji) jest sumą nieskorelowanych błędów odchyłeń od modelu funkcji regresji i błędu estymacji wartości regresyjnej:

$$S(y_{\mathbf{x}_0}^P) = \sqrt{MS_E \left[ 1 + [1 \quad \mathbf{x}_0] \mathbf{V}_0^{-1} [1 \quad \mathbf{x}_0]^T \right]} . \quad (4.36)$$

Podobnie jak w przypadku wartości regresyjnej możemy wyznaczyć przedział ufności dla prawdziwej wartości zmiennej losowej  $Y$  przy ustalonych wartościach  $\mathbf{x}_0$  zmiennych niezależnych:

$$y_{\mathbf{x}_0} \in \langle \hat{m}(\mathbf{x}_0) - t_{\alpha, n-k-1} S(y_{\mathbf{x}_0}^P); \hat{m}(\mathbf{x}_0) + t_{\alpha, n-k-1} S(y_{\mathbf{x}_0}^P) \rangle \text{ z } P = 1 - \alpha . \quad (4.37)$$

*Przykład 4.* Wykorzystując dane empiryczne podane w przykładzie 3 oraz wyniki estymacji modelu funkcji regresji wyznaczmy prognozowaną wartość indywidualnej wydajności pracy dla pracownika w wieku 40 lat ( $x_2$ ) mającego 10-letni staż pracy ( $x_1$ ).

Wektor ustalonych wartości zmiennych niezależnych  $\mathbf{x}_0$  ma postać  $\mathbf{x}_0 = [10 \quad 40]$ , na tej podstawie wyznaczamy ocenę pojedynczej realizacji zmiennej losowej  $Y$ :

$$\hat{y}_{\mathbf{x}_0} = \hat{m}(\mathbf{x}_0) = [1 \quad 10 \quad 40] \cdot \begin{bmatrix} 30,2677 \\ 3,1826 \\ 1,4325 \end{bmatrix} = 119,395 .$$

Do wyznaczenia błędów estymacji wartości regresyjnej i błędu prognozy musimy dysponować macierzą  $\mathbf{V}_0^{-1}$ , powinniśmy więc wyznaczyć brakujący wektor  $\mathbf{D}_2$  (macierz  $\mathbf{V}^{-1}$  oraz liczbę  $A_0$  mamy już wyznaczone w przykładzie 3).

$$\mathbf{D}_2 = -\frac{1}{20} [120 \quad 656] \cdot \begin{bmatrix} 0,00684 & -0,00146 \\ -0,00146 & 0,000791 \end{bmatrix} = [0,0069 \quad -0,01719] .$$

Możemy już zestawić macierz  $\mathbf{V}_0^{-1}$ :

$$\mathbf{V}_0^{-1} = \left[ \begin{array}{c|cc} 0,5724 & 0,0069 & -0,01719 \\ \hline 0,0069 & 0,00684 & -0,00146 \\ -0,01719 & -0,00146 & 0,000719 \end{array} \right] .$$

Znajdujemy kolejno dalsze potrzebne elementy:

$$\begin{aligned}
 [1 \quad \mathbf{x}_0] \cdot \mathbf{V}_0^{-1} &= [1 \mid 10 \quad 40] \cdot \left[ \begin{array}{c|cc} 0,5724 & 0,0069 & -0,01719 \\ \hline 0,0069 & 0,00684 & -0,00146 \\ -0,01719 & -0,00146 & 0,000719 \end{array} \right] = \\
 &= [-0,0462 \quad 0,0168 \quad -0,00015] \\
 ([1 \quad \mathbf{x}_0] \cdot \mathbf{V}_0^{-1}) [1 \quad \mathbf{x}_0]^T &= [-0,0462 \quad 0,0168 \quad -0,00015] \cdot \begin{bmatrix} 1 \\ 10 \\ 40 \end{bmatrix} = 0,11628
 \end{aligned}$$

Możemy już wyznaczyć błąd estymacji wartości regresyjnej i zbudować przedział ufności dla oczekiwanej wartości zmiennej losowej  $Y$  przy ustalonym wektorze  $\mathbf{x}_0$ :

$$S_{\hat{m}(\mathbf{x}_0)} = \sqrt{156,03 \cdot 0,11628} = 4,2595$$

$$\hat{m}(\mathbf{x}_0) \in < 119,395 - 2,11 \cdot 4,2595; 119,395 + 2,11 \cdot 4,2595 > \quad z \quad P = 1 - 0,05$$

$$\hat{m}(\mathbf{x}_0) \in < 110,408; 128,381 > \quad z \quad P = 0,95 \quad .$$

Wyznaczamy dalej błąd prognozy i budujemy przedział ufności dla realizacji pojedynczej wartości zmiennej losowej  $Y$  przy ustalonym wektorze  $\mathbf{x}_0$ :

$$S(y_{\mathbf{x}_0}^P) = \sqrt{156,03 \cdot (1 + 0,11628)} = 13,1974$$

$$y_{\mathbf{x}_0}^P \in < 119,395 - 2,11 \cdot 13,1974; 119,395 + 2,11 \cdot 13,1974 > \quad z \quad P = 1 - 0,05$$

$$y_{\mathbf{x}_0}^P \in < 99,5485; 147,2387 > \quad z \quad P = 0,95 \quad .$$

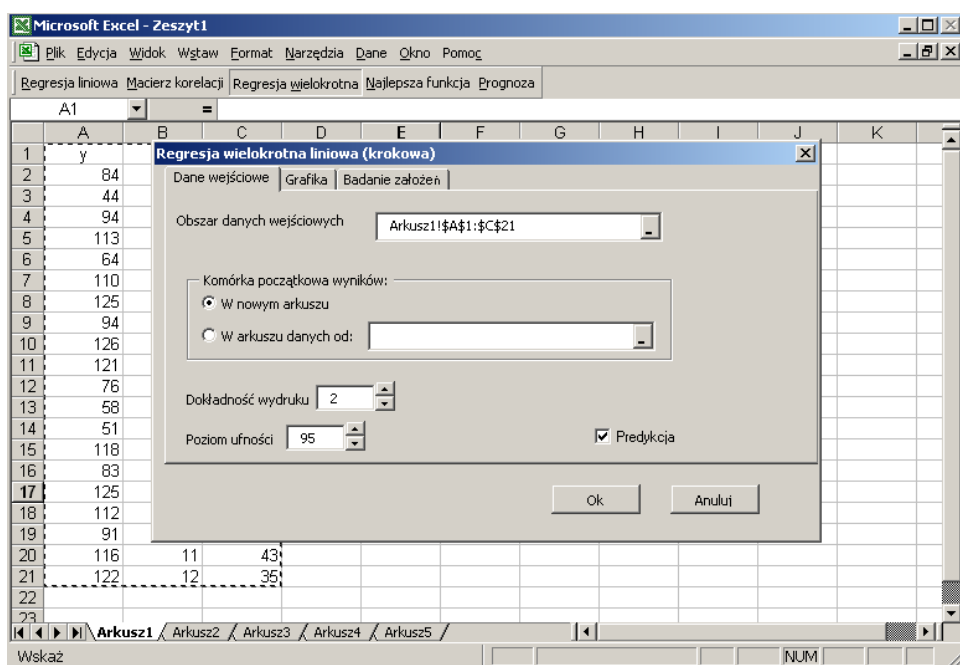
Uzyskane wyniki można zinterpretować następująco: przy ustalonym wektorze zmiennych niezależnych mamy prawo oczekiwać, że średnia indywidualna wydajność pracy będzie równa 119,395 z błędem  $\pm 4,2595$ . Przedział ufności  $<110,408; 128,381>$  pokrywa oczekiwaną wartość zmiennej losowej  $Y$  przy podanym wektorze  $\mathbf{x}_0$  z prawdopodobieństwem  $p = 0,95$ .

Przy ustalonym wektorze  $\mathbf{x}_0$  zmiennych niezależnych ocena realizacji pojedynczej wartości zmiennej losowej  $Y$  jest równa 119,395 z błędem  $\pm 13,1974$ . Przedział ufności o krańcach  $<99,5485; 147,2387>$  pokrywa pojedynczą realizację zmiennej  $Y$  przy ustalonym wektorze  $\mathbf{x}_0$  z prawdopodobieństwem  $p = 0,95$ .

## 4.5. Automatyzacja obliczeń

Jak to wynika z przedstawionej teorii oraz zademonstrowanych przykładów obliczenia związane z estymacją modelu regresji wielokrotnej liniowej, ze zbadaniem istotności modelu, a w szczególności z prognozą są dość skomplikowane. Można temu zaradzić korzystając z wyspecjalizowanego oprogramowania statystycznego lub z przygotowanych w Excelu aplikacji. Pokażę teraz, jak można wykorzystać (pokazaną już w przykładzie 2) aplikację *RegresjaNowa.xls*<sup>5</sup> przygotowaną dla potrzeb zajęć realizowanych z przedmiotów statystycznych.

Po wczytaniu arkusza *RegresjaNowa.xls* zostaje zainstalowany i wyświetlony dedykowany pasek narzędziowy zawierający kilka przycisków pozwalających na uruchomienie określonej analizy.



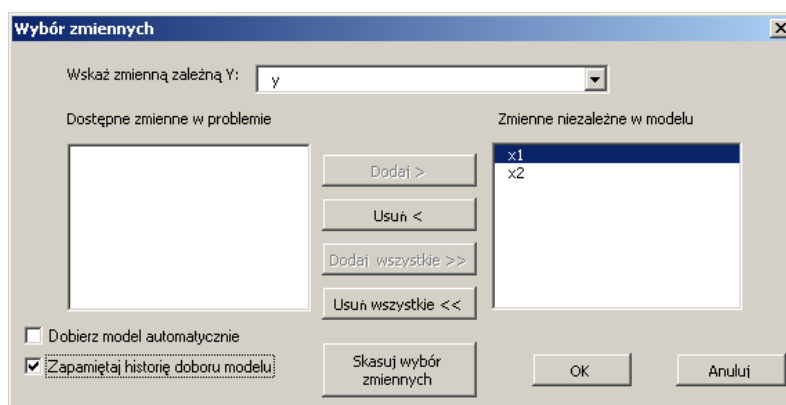
Rysunek 8. Widok arkusza z danymi wyjściowymi i wywołanym oknem dialogowym procedury wykonującej obliczenia regresji wielokrotnej.

<sup>5</sup> do pobrania ze strony <http://www.wszim-sochaczew.edu.pl/download/ekonometria>

Wskazany jest obszar danych wyjściowych (wraz z nazwami – poza oknem dialogowym obszar danych wskazany jest przerywanymi liniami) i zadysponowany nowy arkusz jako miejsce zwrócenia wyników. Ustawione są także takie elementy jak poziom ufności, dokładność wydruku oraz wskazana jest prognoza.

Po zatwierdzeniu ustawień przyciskiem OK wykonywany jest odpowiedni kod programu przygotowujący wyjściową macierz sum kwadratów i iloczynów odchyłeń oraz wektor średnich dla wszystkich zmiennych – bez rozróżniania na tym etapie roli zmiennych w modelu regresji.

Po wykonaniu tych prac przygotowawczych program wyświetla kolejne okno dialogowe w celu ustalenia roli zmiennych.



**Rysunek 9.** Okno dialogowe pozwalające na określenie roli zmiennych w modelu regresji.

W lewej liście program wyświetla wszystkie dostępne zmienne w modelu, spośród nich jedna musi być wskazana jako zmienna zależna, jej wyboru dokonujemy w liście 'Wskaż zmienną zależną'.

Z pozostałych zmiennych wybieramy te z nich, które mają pełnić rolę zmiennych niezależnych, do dyspozycji mamy przyciski 'Dodaj' i 'Dodaj wszystkie'.

Pole wyboru 'Dobierz model automatycznie' pozwala na automatyczny dobór najlepszego, wg programu, modelu funkcji regresji (nie polecam tej opcji).

Po określeniu roli zmiennych potwierdzonej akceptacją przycisku OK program wykonuje wszystkie obliczenia związane z wyestymowaniem modelu funkcji regresji,

zbadaniem jego istotności oraz zbadaniem istotności cząstkowych współczynników regresji.

**Regresja wielokrotna - wyniki doboru modelu**

Badanie istotności modelu

| Zmiennosc | St. swobody | Sr. kw. odchylen | Femp.  | p-value |
|-----------|-------------|------------------|--------|---------|
| Modelu    | 2           | 5 398,032        | 34,596 | 0,000   |
| Odchylen  | 17          | 156,029          |        |         |

Krok 1 Współczynnik determinacji 80,28%

| Zmienna | wsp. regresji | t emp. | p-value |
|---------|---------------|--------|---------|
| x1      | 3,183         | 3,081  | 0,007   |
| x2      | 1,433         | 4,078  | 0,001   |

Buttons: Usun zmienną i przelicz model, Model jest dobrany

**Rysunek 10.** Okno dialogowe wyświetlające wyniki doboru modelu na danym etapie analizy.

W pokazanym oknie znajdziemy wszystkie te elementy, które wyznaczaliśmy rozwiązując przykład 3. Potrzebna jest tu jedna uwaga: prezentowany moduł regresji przeznaczony jest do dobierania najlepszego modelu tzw. metodą regresji krokowej, stąd obecność przycisku pozwalającego na usunięcie wybranej zmiennej i ponowne przeliczenie modelu. Problemowi regresji krokowej będzie poświęcony kolejny rozdział tego skryptu.

Po akceptacji przycisku 'Model jest dobrany' program tworzy nowy arkusz wraz z kompletem wyników. Poza znanymi nam takimi elementami jak oceny parametrów modelu, błędy standardowe ich ocen, dolne i górne granice przedziałów ufności program zwraca także wartości empiryczne statystyki  $t$ -Studenta dla weryfikacji hipotez zerowych o tym, że dany współczynnik regresji jest równy zero (wraz ze stałą regresji). Dla każdej z wyznaczonych statystyk obliczany jest krytyczny poziom istotności ( $p$ -value).

Zwracana jest tabela analizy wariancji z wartością empiryczną testu  $F$  Fishera oraz krytycznym poziomem istotności. W tabeli analizy wariancji znajdziemy tak potrzebną dla wyznaczenia prognozy liczbę stopni swobody dla zmienności resztowej oraz średni kwadrat odchylen od modelu  $MS_E$ .

Program zwraca także dwie miary dobroci dopasowania modelu: współczynnik korelacji wielokrotnej liniowej oraz współczynnik determinacji.

Dodatkowo, na potrzeby przyszłej prognozy, program odtwarza i wyprowadza macierz odwrotną  $V_o^{-1}$ .

|    | A  | B            | C            | D        | E        | F    | G         | H |
|----|--|--------------|--------------|----------|----------|------|-----------|---|
| 1  | Wyniki estymacji modelu regresji wielokrotnej  |              |              |          |          |      |           |   |
| 2  | Zmienna zależna: y   |              |              |          |          |      |           |   |
| 3  | Macierz ocen parametrów, błędy standardowe, dolna i górna granica ufności, t empiryczne, p-value |              |              |          |          |      |           |   |
| 4  |  | b()          | Sb()         | dgu()    | ggu()    | t()  | p-value() |   |
| 5  | Stała  | 30,27        | 9,45         | 10,33    | 50,21    | 3,20 | 0,00522   |   |
| 6  | x1   | 3,18         | 1,03         | 1,00     | 5,36     | 3,08 | 0,00678   |   |
| 7  | x2   | 1,43         | 0,35         | 0,69     | 2,17     | 4,08 | 0,00079   |   |
| 8  |  |              |              |          |          |      |           |   |
| 9  | Badanie istotności regresji testem F Fishera-Snedecora   |              |              |          |          |      |           |   |
| 10 | Zmiennosc  | St. swobod   | Śr. kw. odcl | Femp.    | p-value  |      |           |   |
| 11 | Modelu   | 2            | 5398,03      | 34,60    | 1,02E-06 |      |           |   |
| 12 | Resztowa   | 17           | 156,0286     |          |          |      |           |   |
| 13 |  |              |              |          |          |      |           |   |
| 14 | Wsp. korelacji   | 0,896        |              |          |          |      |           |   |
| 15 | Wsp. determinacji  | 80,3%        |              |          |          |      |           |   |
| 16 |  |              |              |          |          |      |           |   |
| 17 | Elementy macierzy odwrotnej $V_o$ niezbędnej do prognozowania                                    |              |              |          |          |      |           |   |
| 18 |  | 0,572407738  | 0,0069       | -0,01719 |          |      |           |   |
| 19 |  | 0,006900376  | 0,00684      | -0,00146 |          |      |           |   |
| 20 |  | -0,017189329 | -0,00146     | 0,000791 |          |      |           |   |

Rysunek 11. Wyniki estymacji modelu funkcji regresji z przykładu 3.

Przycisk 'Prognoza' pozwala na wyznaczenie prognozowanych wartości zmiennej  $Y$  dla ustalonych wektorów zmiennych niezależnych.

Prognozowanie w regresji wielokrotnej

Wskaż obszar oszacowań współczynników regresji: Arkusz7!\$B\$5:\$B\$7

Wskaż obszar st swobody i MSe: Arkusz7!\$B\$12:\$C\$12

Wskaż obszar macierzy odwrotnej do  $V_o$ : Arkusz7!\$A\$18:\$C\$20

Wskaż obszar zmiennych niezależnych: Arkusz7!\$B\$27:\$C\$47

Poziom ufności: 95

OK

Rysunek 12. Okno dialogowe modułu prognozowania w trakcie określania niezbędnych danych dla prognozy.

Poniżej pokazane są wyniki prognoz dla wyjściowych wektorów zmiennych niezależnych. Ostatni wiersz jest prognozą wykonaną dla wektora  $x_0 = [10 \ 40]$ .

Policzone są te wszystkie elementy, które omawiałem w rozdziale poświęconym prognozowaniu.

|    | B   | C  | D      | E         | F      | G      | H         | I       | J       |
|----|-----|----|--------|-----------|--------|--------|-----------|---------|---------|
|    | x1  | x2 | Y teor | Bł. Stand | Dgufn. | Ggufn. | Bł. Pred. | Dgpred. | Ggpred. |
| 27 | 1   | 19 | 60,67  | 5,1505    | 49,80  | 71,53  | 13,5113   | 32,16   | 89,17   |
| 28 | 0,5 | 21 | 61,94  | 5,2610    | 50,84  | 73,04  | 13,5539   | 33,35   | 90,54   |
| 29 | 2,5 | 25 | 74,04  | 3,9920    | 65,61  | 82,46  | 13,1135   | 46,37   | 101,70  |
| 30 | 9   | 35 | 109,05 | 3,8722    | 100,88 | 117,22 | 13,0776   | 81,46   | 136,64  |
| 31 | 3,5 | 25 | 77,22  | 3,6181    | 69,59  | 84,85  | 13,0046   | 49,78   | 104,66  |
| 32 | 6   | 47 | 116,69 | 5,7185    | 104,63 | 128,76 | 13,7379   | 87,71   | 145,68  |
| 33 | 7,5 | 48 | 122,90 | 5,3229    | 111,67 | 134,13 | 13,5780   | 94,25   | 151,55  |
| 34 | 7   | 28 | 92,66  | 3,7287    | 84,79  | 100,52 | 13,0358   | 65,15   | 120,16  |
| 35 | 7   | 45 | 117,01 | 4,6566    | 107,18 | 126,83 | 13,3309   | 88,88   | 145,14  |
| 36 | 11  | 40 | 122,58 | 4,9461    | 112,14 | 133,01 | 13,4348   | 94,23   | 150,92  |
| 37 | 2   | 23 | 69,58  | 4,3425    | 60,42  | 78,74  | 13,2245   | 41,68   | 97,48   |
| 38 | 2   | 23 | 69,58  | 4,3425    | 60,42  | 78,74  | 13,2245   | 41,68   | 97,48   |
| 39 | 3   | 21 | 69,90  | 4,2959    | 60,83  | 78,96  | 13,2092   | 42,03   | 97,77   |
| 40 | 9,5 | 31 | 104,91 | 4,9141    | 94,54  | 115,28 | 13,4230   | 76,59   | 133,23  |
| 41 | 4   | 25 | 78,81  | 3,5310    | 71,36  | 86,26  | 12,9806   | 51,42   | 106,20  |
| 42 | 5   | 49 | 116,37 | 6,9761    | 101,66 | 131,09 | 14,3071   | 86,19   | 146,56  |
| 43 | 8,5 | 46 | 123,22 | 4,5756    | 113,56 | 132,87 | 13,3028   | 95,15   | 151,28  |
| 44 | 8   | 27 | 94,41  | 4,6385    | 84,62  | 104,19 | 13,3246   | 66,29   | 122,52  |
| 45 | 11  | 43 | 126,87 | 4,9060    | 116,52 | 137,23 | 13,4200   | 98,56   | 155,19  |
| 46 | 12  | 35 | 118,60 | 6,3874    | 105,12 | 132,07 | 14,0295   | 89,00   | 148,20  |
| 47 | 10  | 40 | 119,39 | 4,2595    | 110,41 | 128,38 | 13,1974   | 91,55   | 147,24  |

Rysunek 13. Widok arkusza z wynikami prognoz dla danych wyjściowych oraz dla wektora  $x_0 = [10 \ 40]$ .

## 4.6. Regresja krokowa

W rozdziale 4.1 wstępnie zasygnalizowałem problem doboru zmiennych w modelu regresji w sytuacji, gdy macierz sum kwadratów i iloczynów odchyłeń  $\mathbf{V}$  nie jest macierzą diagonalną. Oznacza to, że zmienne  $X_i$  nie są niezależne, nie są wtedy niezależne także statystyki  $t$ -Studenta wykorzystywane do weryfikacji hipotez o istotności cząstkowych współczynników regresji. Problemy, które mogą się pojawić w takiej sytuacji przedstawię w kolejnym przykładzie liczbowym.

*Przykład 5.* Dane wyjściowe opisujące zależność indywidualnej wydajności pracy (zmienna  $y$ ) od pięciu zmiennych niezależnych (później wyjaśnię jakie są to zmienne).

| $y$ | $x_1$ | $x_2$ | $x_3$   | $x_4$  | $x_5$ |
|-----|-------|-------|---------|--------|-------|
| 84  | 1     | 19    | 0,0000  | 2,9444 | 19    |
| 44  | 0,5   | 21    | -0,6931 | 3,0445 | 10,5  |
| 94  | 2,5   | 25    | 0,9163  | 3,2189 | 62,5  |
| 113 | 9     | 35    | 2,1972  | 3,5553 | 315   |
| 64  | 3,5   | 25    | 1,2528  | 3,2189 | 87,5  |
| 110 | 6     | 47    | 1,7918  | 3,8501 | 282   |
| 125 | 7,5   | 48    | 2,0149  | 3,8712 | 360   |
| 94  | 7     | 28    | 1,9459  | 3,3322 | 196   |
| 126 | 7     | 45    | 1,9459  | 3,8067 | 315   |
| 121 | 11    | 40    | 2,3979  | 3,6889 | 440   |
| 76  | 2     | 23    | 0,6931  | 3,1355 | 46    |
| 58  | 2     | 23    | 0,6931  | 3,1355 | 46    |
| 51  | 3     | 21    | 1,0986  | 3,0445 | 63    |
| 118 | 9,5   | 31    | 2,2513  | 3,4340 | 294,5 |
| 83  | 4     | 25    | 1,3863  | 3,2189 | 100   |
| 125 | 5     | 49    | 1,6094  | 3,8918 | 245   |
| 112 | 8,5   | 46    | 2,1401  | 3,8286 | 391   |
| 91  | 8     | 27    | 2,0794  | 3,2958 | 216   |
| 116 | 11    | 43    | 2,3979  | 3,7612 | 473   |
| 122 | 12    | 35    | 2,4849  | 3,5553 | 420   |

W oparciu o te dane oszacujemy liniową funkcję regresji opisującą zależność między zmienną losową  $Y$  a pozostałymi zmiennymi:

$$m(x_1, x_2, x_3, x_4, x_5) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 .$$

Do estymacji modelu wykorzystamy arkusz *RegresjaNowa.xls* z jego procedurą ‘Regresja wielokrotna’.

Poniżej pokazane jest okno dialogowe z wynikami estymacji wstępnie założonego modelu z 5-cioma zmiennymi niezależnymi.

| Badanie istotności modelu |             |                  |        |         |
|---------------------------|-------------|------------------|--------|---------|
| Zmiennosc                 | St. swobody | Sr. kw. odchylen | Femp.  | p-value |
| Modelu                    | 5           | 2 184,032        | 12,093 | 0,000   |
| Odchylen                  | 14          | 180,599          |        |         |

Krok 1 Współczynnik determinacji 81,2%

| Zmienna | wsp. regresji | t emp.  | p-value |
|---------|---------------|---------|---------|
| x1      | 8,376         | 0,924   | 0,371   |
| x2      | 3,098         | 0,541   | 0,597   |
| x3      | - 1,815       | - 0,134 | 0,895   |
| x4      | - 29,286      | - 0,162 | 0,874   |
| x5      | - 0,147       | - 0,702 | 0,494   |

Rysunek 14. Wyniki estymacji modelu funkcji regresji z przykładu 5.

Analiza statystyki  $F$  Fishera wskazuje, że hipotezę zerową o braku związku między zmienną losową  $Y$  a zmiennymi  $X_i$  ( $i = 1, 2, \dots, 5$ ) musimy odrzucić na rzecz hipotezy alternatywnej. Oznacza to, że **istnieje związek funkcyjny między zmienną  $Y$  a co najmniej jedną zmienną niezależną  $X_i$ .**

Analiza wartości empirycznych statystyk  $t_i$  ( $i = 1, 2, \dots, 5$ ) i odpowiadających im krytycznych poziomów istotności wskazuje jednak, że nie mamy podstaw do odrzucenia żadnej z pięciu hipotez o istotności cząstkowych współczynników regresji. W konsekwencji z modelu funkcji regresji **powinniśmy usunąć wszystkie 5 zmiennych**, co prowadziłoby do wniosku, że **nie istnieje związek funkcyjny między zmienną  $Y$  a badanymi zmiennymi niezależnymi.**

Mamy więc **pozorną** sprzeczność między wnioskiem ogólnym opartym o wyniki weryfikacji hipotezy o istotności regresji wielokrotnej a wnioskami szczegółowymi opartymi o weryfikację 5 hipotez o istotności cząstkowych współczynników regresji.

Taka sytuacja spowodowana jest tym, że zmienne  $X_i$  są dość silnie skorelowane (nie są niezależne w sensie niezależności zmiennych losowych). Miarą stopnia skorelowania tych zmiennych mogą być choćby współczynniki korelacji liniowych. Ponownie skorzystamy z arkusza *RegresjaNowa.xls*, w którym udostępniona jest procedura wyznaczającą macierz ocen współczynników korelacji liniowych między zmiennymi wraz z macierzą krytycznych poziomów istotności dla weryfikacji hipotez zerowych o braku korelacji.

| Regresja liniowa |     | Macierz korelacji | Regresja wielokrotna | Najlepsza funkcja | Prognoza |     |   |   |   |
|------------------|-----|-------------------|----------------------|-------------------|----------|-----|---|---|---|
| A1               | =   |                   |                      |                   |          |     |   |   |   |
|                  | A   | B                 | C                    | D                 | E        | F   | G | H | I |
| 1                | y   | x1                | x2                   | x3                | x4       | x5  |   |   |   |
| 2                | 84  | 1                 | 19                   | 0,0000            | 2,9444   | 19  |   |   |   |
| 3                | 44  |                   |                      |                   |          |     |   |   |   |
| 4                | 94  |                   |                      |                   |          |     |   |   |   |
| 5                | 113 |                   |                      |                   |          |     |   |   |   |
| 6                | 64  |                   |                      |                   |          |     |   |   |   |
| 7                | 110 |                   |                      |                   |          |     |   |   |   |
| 8                | 125 |                   |                      |                   |          |     |   |   |   |
| 9                | 94  |                   |                      |                   |          |     |   |   |   |
| 10               | 126 |                   |                      |                   |          |     |   |   |   |
| 11               | 121 |                   |                      |                   |          |     |   |   |   |
| 12               | 76  |                   |                      |                   |          |     |   |   |   |
| 13               | 58  |                   |                      |                   |          |     |   |   |   |
| 14               | 51  |                   |                      |                   |          |     |   |   |   |
| 15               | 118 |                   |                      |                   |          |     |   |   |   |
| 16               | 83  |                   |                      |                   |          |     |   |   |   |
| 17               | 125 |                   |                      |                   |          |     |   |   |   |
| 18               | 112 |                   |                      |                   |          |     |   |   |   |
| 19               | 91  | 8                 | 27                   | 2,0794            | 3,2958   | 216 |   |   |   |
| 20               | 116 | 11                | 43                   | 2,3979            | 3,7612   | 473 |   |   |   |
| 21               | 122 | 12                | 35                   | 2,4849            | 3,5553   | 420 |   |   |   |
| 22               |     |                   |                      |                   |          |     |   |   |   |

**Wyznaczenie macierzy współczynników korelacji liniowej między zmiennymi**

Wskaż obszar danych wejściowych:

Zapisz wyniki obliczeń:

w nowym arkuszu

w bieżącym arkuszu od komórki

Dokładność wydruku:

Pokaż współczynniki korelacji i krytyczny poziom istotności

OK Anuluj

Rysunek 15. Fragment arkusza z danymi przykładu 5 i otwartym oknem dialogowym procedury 'Macierz korelacji' arkusza RegresjaNowa.xls .

| Regresja liniowa |   | Macierz korelacji | Regresja wielokrotna | Najlepsza funkcja | Prognoza |        |   |   |
|------------------|---|-------------------|----------------------|-------------------|----------|--------|---|---|
| L1               | =   |                   |                      |                   |          |        |   |   |
|                  | A   | B                 | C                    | D                 | E        | F      | G | H |
| 1                | Macierz współczynników korelacji liniowych między zmiennymi |                   |                      |                   |          |        |   |   |
| 2                | y   | x1                | x2                   | x3                | x4       |        |   |   |
| 3                | x1  | 0,781             |                      |                   |          |        |   |   |
| 4                | x2  | 0,832             | 0,628                |                   |          |        |   |   |
| 5                | x3  | 0,790             | 0,927                | 0,669             |          |        |   |   |
| 6                | x4  | 0,859             | 0,692                | 0,994             | 0,732    |        |   |   |
| 7                | x5  | 0,854             | 0,946                | 0,822             | 0,876    | 0,859  |   |   |
| 8                |   |                   |                      |                   |          |        |   |   |
| 9                | Macierz krytycznych poziomów istotności                     |                   |                      |                   |          |        |   |   |
| 10               | y   | x1                | x2                   | x3                | x4       |        |   |   |
| 11               | x1  | 0,0000            |                      |                   |          |        |   |   |
| 12               | x2  | 0,0000            | 0,0030               |                   |          |        |   |   |
| 13               | x3  | 0,0000            | 0,0000               | 0,0013            |          |        |   |   |
| 14               | x4  | 0,0000            | 0,0007               | 0,0000            | 0,0002   |        |   |   |
| 15               | x5  | 0,0000            | 0,0000               | 0,0000            | 0,0000   | 0,0000 |   |   |
| 16               |   |                   |                      |                   |          |        |   |   |

Rysunek 16. Macierz współczynników korelacji liniowych między zmiennymi.

Analiza wyników przedstawionych na rys. 16 wskazuje na istnienie **bardzo silnych** związków liniowych między wszystkimi parami zmiennych. Tak jak korelacja między zmienną  $y$  a pozostałymi zmiennymi jest naturalna i wręcz oczekiwana, to między zmiennymi  $x_i$  jest szkodliwa i jest zapowiedzią przyszłych problemów przy doborze modelu funkcji regresji.

Konsekwencją tego, że zmienne niezależne są skorelowane jest **niemożność określenia w jednym kroku**, w wyniku zweryfikowania serii hipotez o istotności cząstkowych współczynników regresji, zestawu tych zmiennych niezależnych, które powinny pozostać w modelu funkcji regresji. Oznacza to konieczność wypracowania innej metody pozwalającej na określenie najlepszego zestawu zmiennych niezależnych.

Jedną z takich metod jest regresja **krokowa**. W teorii statystyki znane są dwie wersje tej metody: jedna z nich polega na dodawaniu zmiennych niezależnych, a druga na usuwaniu zmiennych (regresja krokowa wsteczna). Ja zaproponuję Czytelnikom tego skryptu regresję krokową **wsteczną**.

Metodę doboru modelu funkcji regresji metodą regresji krokowej wstecznej można przedstawić w kilku punktach:

1. Określamy wyjściowy, maksymalny zestaw zmiennych niezależnych w modelu funkcji regresji i estymujemy ten model (krok 1).
2. Z modelu funkcji regresji eliminujemy tę zmienną niezależną, dla której wartość bezwzględna statystyki  $t$ -Studenta dla weryfikacji hipotez o istotności cząstkowych współczynników regresji jest najmniejsza (tym samym krytyczny poziom istotności jest największy).
3. Ponownie estymujemy model funkcji regresji i przechodzimy do p. 2.
4. Krok 2 i 3 trwają tak długo, dopóki w modelu funkcji regresji nie pozostaną tylko istotne zmienne niezależne.

W trakcie wykonywania regresji krokowej powinniśmy obserwować zmianę średniego kwadratu odchyleń od modelu funkcji regresji -  $MS_E$  oraz współczynnika determinacji  $R^2$ .

W regresji krokowej wstecznej w każdym kroku zmniejszamy liczbę zmiennych w modelu, co w konsekwencji **musi** zmniejszać wartość współczynnika determinacji. W sytuacji, gdy z modelu usuwamy zmienną nieistotną, to zmniejszenie współczynnika determinacji jest minimalne (nieznaczące).

Usunięcie nieistotnej zmiennej niezależnej z modelu funkcji regresji powoduje zwiększenie o jeden liczby stopni swobody dla zmienności resztowej, co w połączeniu z faktem, że nastąpiło nieznaczne zwiększenie sumy kwadratów odchyleń dla zmienności resztowej powoduje zmniejszenie średniego kwadratu odchyleń od modelu funkcji regresji, a o to także chodzi w regresji krokowej.

Reasumując, celem regresji krokowej jest pozostawienie w modelu funkcji regresji minimalnego zestawu zmiennych niezależnych przy jednoczesnej maksymalizacji współczynnika determinacji i minimalizacji średniego kwadratu odchyleń od modelu regresji.

Poniżej pokazuję kolejne kroki doboru modelu funkcji regresji metodą regresji krokowej.

**Regresja wielokrotna - wyniki doboru modelu**

Badanie istotności modelu

| Zmienność | St. swobody | Sr. kw. odchyłeń | Femp.  | p-value |
|-----------|-------------|------------------|--------|---------|
| Modelu    | 5           | 2 184,032        | 12,093 | 0,000   |
| Odchyłeń  | 14          | 180,599          |        |         |

Krok 1 Współczynnik. determinacji 81,2%

| Zmienna | wsp. regresji | t emp.  | p-value |
|---------|---------------|---------|---------|
| x1      | 8,376         | 0,924   | 0,371   |
| x2      | 3,098         | 0,541   | 0,597   |
| x3      | - 1,815       | - 0,134 | 0,895   |
| x4      | - 29,286      | - 0,162 | 0,874   |
| x5      | - 0,147       | - 0,702 | 0,494   |

Buttons: Usun zmienną i przelicz model, Model jest dobrany

Rysunek 17. Krok 1 regresji krokowej, wyniki estymacji modelu z pięcioma zmiennymi, wskazana jest zmienna x3 do (potencjalnego) usunięcia.

**Regresja wielokrotna - wyniki doboru modelu**

Badanie istotności modelu

| Zmienność | St. swobody | Sr. kw. odchyłeń | Femp.  | p-value |
|-----------|-------------|------------------|--------|---------|
| Modelu    | 4           | 2 729,229        | 16,171 | 0,000   |
| Odchyłeń  | 15          | 168,776          |        |         |

Krok 2 Współczynnik. determinacji 81,18% Poprzedni wsp. determinacji 81,2%  
Poprzedni MSE 180,599

| Zmienna | wsp. regresji | t emp.  | p-value |
|---------|---------------|---------|---------|
| x1      | 7,555         | 1,171   | 0,260   |
| x2      | 3,068         | 0,555   | 0,587   |
| x4      | - 32,011      | - 0,184 | 0,856   |
| x5      | - 0,131       | - 0,779 | 0,448   |

Buttons: Usun zmienną i przelicz model, Model jest dobrany

Rysunek 18. Krok 2 regresji krokowej, wyniki estymacji modelu z czterema zmiennymi, wskazana jest zmienna x4 do (potencjalnego) usunięcia.

| Badanie istotności modelu |             |                  |        |         |
|---------------------------|-------------|------------------|--------|---------|
| Zmienność                 | St. swobody | Sr. kw. odchyłeń | Femp.  | p-value |
| Modelu                    | 3           | 3 637,068        | 22,935 | 0,000   |
| Odchyłeń                  | 16          | 158,584          |        |         |

Krok 3 Współczynnik determinacji 81,13% Poprzedni wsp. determinacji 81,18%  
Poprzedni MSE 168,776

| Zmienna | wsp. regresji | t emp.  | p-value |
|---------|---------------|---------|---------|
| x1      | 6,682         | 1,577   | 0,134   |
| x2      | 2,062         | 2,516   | 0,023   |
| x5      | - 0,113       | - 0,849 | 0,408   |

Usun zmienną i przelicz model

Model jest dobrany

Rysunek 19. Krok 3 regresji krokowej, wyniki estymacji modelu z trzema zmiennymi, wskazana jest zmienna x5 do (potencjalnego) usunięcia.

| Badanie istotności modelu |             |                  |        |         |
|---------------------------|-------------|------------------|--------|---------|
| Zmienność                 | St. swobody | Sr. kw. odchyłeń | Femp.  | p-value |
| Modelu                    | 2           | 5 398,032        | 34,596 | 0,000   |
| Odchyłeń                  | 17          | 156,029          |        |         |

Krok 4 Współczynnik determinacji 80,28% Poprzedni wsp. determinacji 81,13%  
Poprzedni MSE 158,584

| Zmienna | wsp. regresji | t emp. | p-value |
|---------|---------------|--------|---------|
| x1      | 3,183         | 3,081  | 0,007   |
| x2      | 1,433         | 4,078  | 0,001   |

Usun zmienną i przelicz model

Model jest dobrany

Rysunek 20. Krok 4 regresji krokowej, wyniki estymacji modelu z dwiema zmiennymi, model jest dobrany (zostały tylko zmienne istotne).

Po zatwierdzeniu modelu do nowego arkusza zwracane są wszystkie wyniki estymacji modelu metodą regresji krokowej wstecznej wraz z historią przebiegu regresji krokowej. Podawane są kolejne kroki, liczba zmiennych w modelu funkcji regresji, liczba

stopni swobody zmienności resztowej, średni kwadrat odchyłeń  $MS_E$ , współczynnik determinacji oraz nazwa usuniętej zmiennej w danym kroku.

| Microsoft Excel - PrzykładRegresjaNowa.xls   |  |             |               |         |            |                  |            |  |
|--|--|-------------|---------------|---------|------------|------------------|------------|--|
| Regresja liniowa Macierz korelacji Regresja wielokrotna Najlepsza funkcja Prognoza |  |             |               |         |            |                  |            |  |
| G30 =  |  |             |               |         |            |                  |            |  |
|  | A  | B           | C             | D       | E          | F                | G          |  |
| 1  | Wyniki estymacji modelu regresji wielokrotnej  |             |               |         |            |                  |            |  |
| 2  | Zmienna zależna: y   |             |               |         |            |                  |            |  |
| 3  | Macierz ocen parametrów, błędy standardowe, dolna i górna granica ufności, t empiryczne, p-value |             |               |         |            |                  |            |  |
| 4  |  | b(i)        | Sb(i)         | dgu(i)  | ggu(i)     | t(i)             | p-value(i) |  |
| 5  | Stała  | 30,27       | 9,45          | 10,33   | 50,21      | 3,20             | 0,01       |  |
| 6  | x1   | 3,18        | 1,03          | 1,00    | 5,36       | 3,08             | 0,01       |  |
| 7  | x2   | 1,43        | 0,35          | 0,69    | 2,17       | 4,08             | 0,00       |  |
| 8  |  |             |               |         |            |                  |            |  |
| 9  | Badanie istotności regresji testem F Fishera-Snedecora   |             |               |         |            |                  |            |  |
| 10   | Zmiennosc  | St. swobody | Sr. kw. odch. | Femp.   | p-value    |                  |            |  |
| 11   | Modelu   | 2           | 5398,03       | 34,60   | 1,02E-06   |                  |            |  |
| 12   | Resztowa   | 17          | 156,028584    |         |            |                  |            |  |
| 13   |  |             |               |         |            |                  |            |  |
| 14   | Wsp. korelacji   | 0,896       |               |         |            |                  |            |  |
| 15   | Wsp. determinacji  | 80,3%       |               |         |            |                  |            |  |
| 16   |  |             |               |         |            |                  |            |  |
| 17   | Elementy macierzy odwrotnej $V_0$ niezbędnej do prognozowania                                    |             |               |         |            |                  |            |  |
| 18   | 0,572407738  | 0,00690038  | -0,01718933   |         |            |                  |            |  |
| 19   | 0,006900376  | 0,00684008  | -0,00146161   |         |            |                  |            |  |
| 20   | -0,017189329   | -0,0014616  | 0,00079143    |         |            |                  |            |  |
| 21   |  |             |               |         |            |                  |            |  |
| 22   | Historia doboru modelu metodą regresji krokowej wstecznej  |             |               |         |            |                  |            |  |
| 23   | Krok   | Zmiennych   | St. sw. błędu | MSe     | Wsp. deter | Zmienna usunięta |            |  |
| 24   | 1  | 5           | 14            | 180,599 | 0,81200    |                  |            |  |
| 25   | 2  | 4           | 15            | 168,776 | 0,81175    | x3               |            |  |
| 26   | 3  | 3           | 16            | 158,584 | 0,81133    | x4               |            |  |
| 27   | 4  | 2           | 17            | 156,029 | 0,80277    | x5               |            |  |

Rysunek 21. Wyniki estymacji modelu funkcji regresji z przykładu 5 wraz z historią doboru modelu metodą regresji krokowej wstecznej.

## 4.7. Uzupełnienie korelacji cząstkowych

W rozdziale 3.3 przedstawiłem problem oszacowania współczynnika korelacji liniowej między dwoma wybranymi zmiennymi losowymi i przy wyeliminowaniu wpływów trzeciej z rozpatrywanych zmiennych. W tym rozdziale możemy tę problematykę rozszerzyć na większą niż trzy liczbę zmiennych losowych. W przykładzie 2 zajmowałem się oszacowaniem współczynników korelacji liniowej między zmiennymi losowymi reprezentującymi średnie ocen studentów uzyskanych w semestrach 1-6. Okazało się wtedy, że korelacja liniowa między ocenami uzyskanymi w semestrze 1 i 6 nie istnieje (nie jest istotna statystycznie). Wyznaczony współczynnik korelacji  $r_{16}$  nie uwzględniał jednak wpływu pozostałych semestrów, a przecież trudno taki wpływ zbagatelizować. Jeżeli chcielibyśmy wyznaczyć faktyczną miarę związku między ocenami uzyskanymi w semestrze 6 i 1, to powinniśmy wyznaczyć współczynnik korelacji cząstkowej postaci  $r_{16,2345}$ . Nie można niestety wyznaczyć jego wartości z wzoru (3.7), ale możemy ten wzór adaptować do obecnej sytuacji:

$$r_{16,2345} = \frac{r_{16} - R_{1,2345} \cdot R_{6,2345}}{\sqrt{(1 - R_{1,2345}^2) \cdot (1 - R_{6,2345}^2)}} \quad (4.38)$$

gdzie  $R_{1,2345}$  jest współczynnikiem korelacji wielokrotnej liniowej wyznaczonej między zmienną o numerze 1 a zmiennymi niezależnymi o numerach 2, 3, 4 i 5. Podobnie  $R_{6,2345}$  jest współczynnikiem korelacji wielokrotnej liniowej wyznaczonej między zmienną o numerze 6, a zmiennymi niezależnymi o numerach 2, 3, 4 i 5.

Korzystając z procedury 'Regresja wielokrotna' i danych empirycznych zapisanych w arkuszu *Wyniki* skoroszytu *MacierzKorelacji.xls* możemy wyznaczyć potrzebne współczynniki korelacji wielokrotnej liniowej otrzymując odpowiednio:

$$R_{1,2345} = 0,588$$

$$R_{6,2345} = 0,668 .$$

Współczynnik korelacji liniowej między semestrem 1 i 6 jest równy 0,035 ( $r_{16}$ ), czyli możemy już wyznaczyć współczynnik korelacji cząstkowej zmiennych 1 i 6 z wyeliminowaniem wpływu zmiennych o numerach 2, 3, 4 i 5 (korzystamy z (4.38)):

$$r_{16,2345} = \frac{0,035 - 0,588 \cdot 0,668}{\sqrt{(1 - 0,588^2) \cdot (1 - 0,668^2)}} = -0,594 .$$

Jak widzimy, mamy zupełnie inny obraz sytuacji: cząstkowy współczynnik korelacji jest nie tylko istotnie mniejszy niż całkowity  $r_{16}$ , ale także zmienił znak z dodatniego na ujemny. Taki wynik można zinterpretować merytorycznie tylko w jeden sposób: ci studenci, którzy byli dobrzy w pierwszym semestrze nie uzyskują tak dobrych ocen w semestrze szóstym. Oznacza to jednocześnie, że studenci ze słabymi ocenami w pierwszym semestrze uzyskują zdecydowanie lepsze oceny w semestrze szóstym.