

## 7.2. Regresja wielokrotna liniowa

W poprzednim rozdziale przedstawiłem taką sytuacją, w której w populacji generalnej  $\pi$  rozważane były tylko dwie zmienne: zmienna losową  $Y$  i zmienna losowa lub rzeczywista  $X$ . Obecny rozdział poświęcony jest takiej sytuacji, w której w populacji generalnej  $\pi$  obserwowane 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:

$$Y \sim N(m(x_1, x_2, \dots, x_k); \sigma_{y/x_1, x_2, \dots, x_k}) . \quad (7.39)$$

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

$$m(x_1, x_2, \dots, x_k) = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k = b_0 + \sum_i^k b_ix_i . \quad (7.40)$$

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 7.40 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 7.40 dowolną obserwację empiryczną możemy przedstawić jako:

$$y_j = b_0 + b_1x_{1j} + b_2x_{2j} + \dots + b_kx_{kj} + e_j \quad \text{dla } j=1, 2, \dots, n . \quad (7.41)$$

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

$$e_j = y_j - (b_0 + b_1x_{1j} + b_2x_{2j} + \dots + b_kx_{kj}) = y_j - \hat{y}_j . \quad (7.42)$$

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  $s$ :

$$s = \sum_{j=1}^n e_j^2 = \sum_{j=1}^n [y_j - (b_0 + b_1x_{1j} + b_2x_{2j} + \dots + b_kx_{kj})]^2 = \text{minimum} \quad (7.43)$$

Wyznaczenie minimum funkcji  $s$  określonej wzorem 7.43 wymaga rozwiązania układu  $k+1$  równań. Problem można trochę uprościć przekształcając równość 7.41 w taki sposób, aby wyeliminować stałą regresji  $b_0$ :

$$y_j - \bar{y} = b_1(x_{1j} - \bar{x}_1) + b_2(x_{2j} - \bar{x}_2) + \dots + b_k(x_{kj} - \bar{x}_k) + e_j \quad (7.42)$$

gdzie

$$b_0 = \bar{y} - (b_1\bar{x}_1 + b_2\bar{x}_2 + \dots + b_k\bar{x}_k) = \bar{y} - \sum_{i=1}^k b_i\bar{x}_i. \quad (7.43)$$

Uwzględniając wzór 7.42 kryterium estymacji można zapisać następująco:

$$s = \sum_{j=1}^n e_j^2 = \sum_{j=1}^n [y_j - \bar{y} - (b_1(x_{1j} - \bar{x}_1) + \dots + b_k(x_{kj} - \bar{x}_k))]^2 = \text{minimum} \quad (7.44)$$

Znalezienie minimum funkcji  $s$  określonej wzorem 7.44 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 (tzw. normalnych).

$$\frac{\partial s}{\partial b_i} = -2 \sum_{j=1}^n [y_j - (b_1(x_{1j} - \bar{x}_1) + \dots + b_k(x_{kj} - \bar{x}_k))] (x_{ij} - \bar{x}_i) = 0 \quad (7.45)$$

Otrzymany w wyniku zastosowania wzoru 7.45 układ równań normalnych możemy zapisać w tradycyjnej postaci:

$$\begin{cases} \hat{b}_1 \text{var} x_1 + \hat{b}_2 \text{cov} x_1 x_2 + \dots + \hat{b}_k \text{cov} x_1 x_k & = & \text{cov} x_1 y \\ \hat{b}_1 \text{cov} x_2 x_1 + \hat{b}_2 \text{var} x_2 + \dots + \hat{b}_k \text{cov} x_2 x_k & = & \text{cov} x_2 y \\ \vdots & & \vdots \\ \hat{b}_1 \text{cov} x_k x_1 + \hat{b}_2 \text{cov} x_k x_2 + \dots + \hat{b}_k \text{var} x_k & = & \text{cov} x_k y \end{cases} \quad (7.46)$$

W układzie równań określonym wzorem 7.46 w miejsce parametrów modelu w populacji generalnej  $b_i$  wprowadzono już ich oceny z próby  $\hat{b}_i$ .

Układ równań normalnych przedstawiony wzorem 7.46 wygodniej jest zapisać w notacji macierzowej:

$$\mathbf{V}\hat{\mathbf{B}} = \mathbf{C} \quad (7.47)$$

gdzie

$$\mathbf{V} = \begin{matrix} (k \times k) \\ \begin{bmatrix} \text{var} x_1 & \text{cov} x_1 x_2 & \dots & \text{cov} x_1 x_k \\ \text{cov} x_2 x_1 & \text{var} x_2 & \dots & \text{cov} x_2 x_k \\ \vdots & \vdots & \dots & \vdots \\ \text{cov} x_k x_1 & \text{cov} x_k x_2 & \dots & \text{var} x_k \end{bmatrix} \end{matrix} \quad (7.48)$$

$$\hat{\mathbf{B}}_{(k \times 1)} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \vdots \\ \hat{b}_k \end{bmatrix} \quad \mathbf{C}_{(k \times 1)} = \begin{bmatrix} \text{cov } x_1 y \\ \text{cov } x_2 y \\ \vdots \\ \text{cov } x_k y \end{bmatrix} \quad (7.49)$$

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 7.47 lewostronnie przez macierz odwrotną do  $\mathbf{V}$  otrzymujemy:

$$\mathbf{V}\hat{\mathbf{B}} = \mathbf{C} \mid \cdot \mathbf{V}^{-1} \Rightarrow \mathbf{V}^{-1}\mathbf{V}\hat{\mathbf{B}} = \mathbf{V}^{-1}\mathbf{C} \Rightarrow \mathbf{I}\hat{\mathbf{B}} = \mathbf{V}^{-1}\mathbf{C} \Rightarrow \hat{\mathbf{B}} = \mathbf{V}^{-1}\mathbf{C} \quad (7.50)$$

Po wyestymowaniu parametrów  $\hat{b}_i$  (dla  $i = 1, 2, \dots, k$ ) z równania 7.50 wyznaczamy ocenę parametru  $b_0$  z wzoru:

$$\hat{b}_0 = \bar{y} - \sum_{i=1}^k \hat{b}_i \bar{x}_i . \quad (7.51)$$

Ostatecznie ocena funkcji regresji z próby ma postać:

$$\hat{m}(x_1, x_2, \dots, x_k) = \hat{b}_0 + \sum_{i=1}^k \hat{b}_i x_i . \quad (7.52)$$

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.

### 7.2.1. 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 częstkowe współczynniki regresji będą jednocześnie równe zero. Tym samym hipotezie o **istotności regresji wielokrotnej liniowej** możemy nadać postać:

$$H_0 : \wedge_i b_i = 0 \text{ (dla } i = 1, 2, \dots, k) \quad (7.53)$$

Weryfikację tak sformułowanej hipotezy zerowej wobec alternatywy  $H_1 : \vee_i b_i \neq 0$  przeprowadzamy testem  $F$  Fishera w analizie wariancji.

Tabela analizy wariancji dla weryfikacji hipotezy o istotności regresji wielokrotnej.

Zmiennosc	Stopnie swobody	Suma kwadratow odchylen	Średni kwadrat odchylen	F empiryczne
Modelu	$v_R = k$	$\text{var } R = \sum_{i=1}^k \hat{b}_i \text{cov } x_i y$	$s_R^2 = \frac{\text{var } R}{v_R}$	$F_R = \frac{s_R^2}{s_E^2}$
Resztowa	$v_E = n - k - 1$	$\text{var } E = \text{var } T - \text{var } R$	$s_E^2 = \frac{\text{var } E}{v_e}$	
Całkowita	$v_T = n - 1$	$\text{var } y = \sum_{j=1}^n (y_j - \bar{y})^2$		

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:

$$H_{0i} : b_i = 0 \text{ wobec } H_{1i} : b_i \neq 0 \text{ dla } i = 1, 2, \dots, k \quad (7.54)$$

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

$$t_i = \frac{\hat{b}_i}{S_{\hat{b}_i}} = \frac{\hat{b}_i}{\sqrt{S_{y/x_1, x_2, \dots, x_k}^2 \cdot v^{ii}}} \quad (7.55)$$

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  $\mathbf{V}$ :

$$S_{\hat{b}_i} = \sqrt{S_{y/x_1, x_2, \dots, x_k}^2 \cdot v^{ii}} \quad (7.56)$$

Statystyka określona wzorem 7.56 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.

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 7.55 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 dziać 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 7.54 nie wchodzi w rachubę.

Działanie takie, jak przedstawione powyżej jest poprawne wtedy tylko, gdy zmienne niezależne (objaśniające)  $X$  są **wzajemnie niezależne**, czyli wtedy, gdy macierz  $\mathbf{V}$  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 przedstawię jeszcze miary dobroci 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  $Y$ :

$$R^2 = \frac{\sum_{i=1}^k \hat{b}_i \text{cov } x_i y}{\text{var } y} \quad (7.57)$$

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  $<0;1>$ , z tym, że najczęściej wyrażamy go w procentach  $<0\%; 100\%>$ .

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

$$R = \sqrt{R^2} = \sqrt{\frac{\sum_{i=1}^k \hat{b}_i \text{cov } x_i, y}{\text{var } y}} \quad (7.58)$$

Współczynnik korelacji wielokrotnej  $R$  przyjmuje swoje wartości z przedziału  $<0; 1>$ , 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)  $S_{y/x_1, x_2, \dots, x_k}^2$ . Im ten średni kwadrat odchyłeń jest mniejszy, tym model jest lepiej dopasowany.

Wielkość  $S_{y/x_1, x_2, \dots, x_k}^2$  wpływa także na błędy estymacji parametrów modelu oraz błąd wartości regresyjnej i błąd predykcji.

### 7.2.2. Regresja krokowa

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 proponuję 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 odchyłeń od modelu funkcji regresji -  $s_{y/x_1, \dots, x_k}^2$  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 odchyłeń dla zmienności resztowej powoduje zmniejszenie średniego kwadratu odchyłeń 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 odchyłeń od modelu regresji.

### 7.2.3. Dokładność oceny parametrów modelu

Parametry modelu 7.41 są szacowane z próby losowej, tym samym ich oceny obciążone są pewnym błędem. Ocenę błędu standardowego cząstkowego,  $i$ -tego współczynnika regresji znajdziemy z wzoru:

$$S_{\hat{b}_i} = \sqrt{S_{y/x_1, x_2, \dots, x_k}^2 \cdot v^{ii}} \quad \text{dla } i = 1, 2, \dots, k \quad (7.59)$$

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

$$b_i \in \langle \hat{b}_i - t_{\alpha, n-k-1} S_{\hat{b}_i}; \hat{b}_i + t_{\alpha, n-k-1} S_{\hat{b}_i} \rangle \quad \text{z } P = 1 - \alpha \quad (7.60)$$

Wyznaczenie oceny błędu standardowego stałej regresji jest trochę bardziej skomplikowane:

$$S_{\hat{b}_0} = \sqrt{S_{y/x_1, x_2, \dots, x_k}^2 \cdot A_0} \quad (7.61)$$

gdzie wielkość  $A_0$  wyznaczana jest z wzoru:

$$A_0 = \frac{1}{n} \left[ \mathbf{1} + \frac{1}{n} \mathbf{D}_1^T \mathbf{V}^{-1} \mathbf{D}_1 \right] . \quad (7.62)$$

Wektor  $\mathbf{D}_1^T$  występujący we wzorze 7.62 jest wektorem sum zmiennych niezależnych wyznaczonych na podstawie  $n$ -elementowej próby losowej:

$$\mathbf{D}_1^T = \left[ \sum_{j=1}^n x_{1j} \quad \sum_{j=1}^n x_{2j} \quad \dots \quad \sum_{j=1}^n x_{kj} \right] . \quad (7.63)$$

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

$$b_0 \in \left( \hat{b}_0 - t_{\alpha, n-k-1} S_{\hat{b}_0}; \hat{b}_0 + t_{\alpha, n-k-1} S_{\hat{b}_0} \right) \text{ z } P = 1 - \alpha . \quad (7.64)$$

Interpretacja zbudowanych zgodnie ze wzorami 7.60 i 7.64 przedziałów ufności dla cząstkowych współczynników regresji i stałej regresji jest standardowa: zbudowany przedział liczbowy pokrywa nieznaną wartość parametru z prawdopodobieństwem  $1 - \alpha$ .

#### 7.2.4. Predykcja w regresji wielokrotnej

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  $\mathbf{x}_0 = [x_{10} \quad x_{20} \quad \dots \quad x_{k0}]$  znajdziemy z wzoru:

$$\hat{m}(\mathbf{x}_0) = \begin{bmatrix} \mathbf{1} & \mathbf{x}_0 \end{bmatrix} \cdot \begin{bmatrix} \hat{b}_0 \\ \hat{\mathbf{B}} \end{bmatrix} = \hat{b}_0 + \sum_{i=1}^k \hat{b}_i x_{i0} . \quad (7.65)$$

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:

$$S_{\hat{m}(\mathbf{x}_0)} = \sqrt{s_{y/x_1, \dots, x_k}^2 \begin{bmatrix} \mathbf{1} & \mathbf{x}_0 \end{bmatrix} \mathbf{V}_0^{-1} \begin{bmatrix} \mathbf{1} \\ \mathbf{x}_0 \end{bmatrix}^T} \quad (7.66)$$

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

$$\mathbf{V}_0^{-1} = \begin{bmatrix} A_0 & \mathbf{D}_2 \\ \mathbf{D}_2^T & \mathbf{V}^{-1} \end{bmatrix} . \quad (7.67)$$

Wyrażenie (liczba)  $A_0$  zostało już wcześniej określone wzorem 7.62, a wektor  $\mathbf{D}_2$  znajdziemy z wzoru:

$$\mathbf{D}_2 = -\frac{1}{n} \mathbf{D}_1 \mathbf{V}^{-1} \quad (7.68)$$

gdzie wektor  $\mathbf{D}_1$  był już określony wzorem 7.63 jako wektor sum obserwacji dla zmiennych niezależnych.

Wykorzystując wartość regresyjną określoną wzorem 7.65 oraz błąd standardowy tej oceny dany wzorem 7.66 budujemy przedział ufności dla wartości regresyjnej:

$$m(\mathbf{x}_0) \in \langle \hat{m}(\mathbf{x}_0) - t_{\alpha, n-k-1} S_{\hat{m}(\mathbf{x}_0)}; \hat{m}(\mathbf{x}_0) + t_{\alpha, n-k-1} S_{\hat{m}(\mathbf{x}_0)} \rangle \quad z \quad P = 1 - \alpha \quad . \quad (7.69)$$

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 wyznaczymy z wzoru:

$$y_{\mathbf{x}_0} = [\mathbf{1} \quad \mathbf{x}_0] \cdot \begin{bmatrix} b_0 \\ \mathbf{B} \end{bmatrix} + e \quad (7.70)$$

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:

$$S(y_{\mathbf{x}_0}^P) = \sqrt{s_{y/x_1, \dots, x_k}^2 \left[ \mathbf{1} + [\mathbf{1} \quad \mathbf{x}_0] \mathbf{V}_0^{-1} [\mathbf{1} \quad \mathbf{x}_0]^T \right]} \quad (7.71)$$

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

$$y_{\mathbf{x}_0} \in \langle \hat{m}(\mathbf{x}_0) - t_{\alpha, n-k-1} S(y_{\mathbf{x}_0}^P); \hat{m}(\mathbf{x}_0) + t_{\alpha, n-k-1} S(y_{\mathbf{x}_0}^P) \rangle \quad z \quad P = 1 - \alpha \quad . \quad (7.72)$$

**Przykład 7.2.** Na podstawie wyników opisujących wielkość zbioru zbóż w tys. ton w województwach dawnego podziału terytorialnego kraju chcemy poszukać związku funkcyjnego między tą zmienną (zbiorami zbóż), a zmiennymi potencjalnie ją kształtującymi: wielkością użytków rolnych w tys. ha, liczbą ciągników w tys. ha oraz wielkością zużycia nawozów NPK w kg/ha. Dane empiryczne zapisałem w arkuszu *Przyklad7.2* skoroszytu *DaneDoRegresji*.

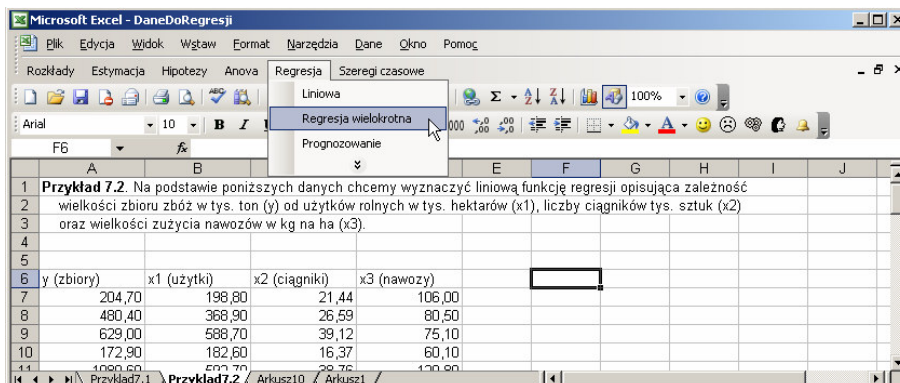
Po wyestymowaniu modelu funkcji regresji wyznaczymy prognozowaną wielkość zbioru zbóż (zmienna  $y$ -ek) dla następujących wartości zmiennych niezależnych:

$x_1$ (wielkość użytków rolnych tys. ha)	450
$x_2$ (liczba ciągników w tys. sztuk)	50
$x_3$ (wielkość nawożenia NPK w kg/ha)	80

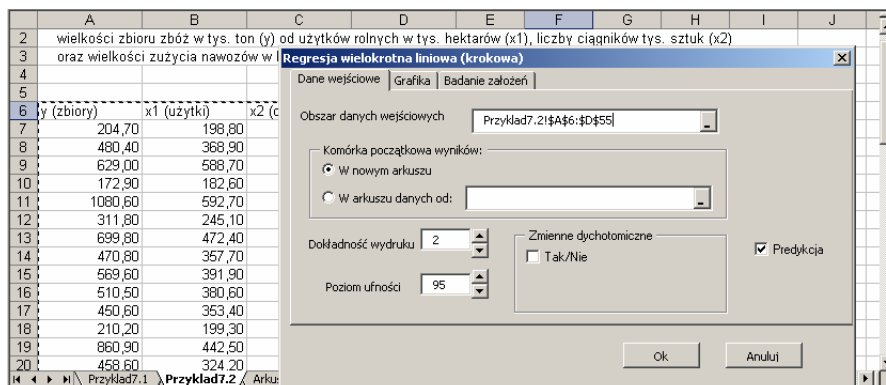
Założymy, że między zmienną zależną (objaśnianą) a zmiennymi niezależnymi zachodzi liniowa zależność:

$$y = m(x_1, x_2, x_3) = b_0 + b_1x_1 + b_2x_2 + b_3x_3.$$

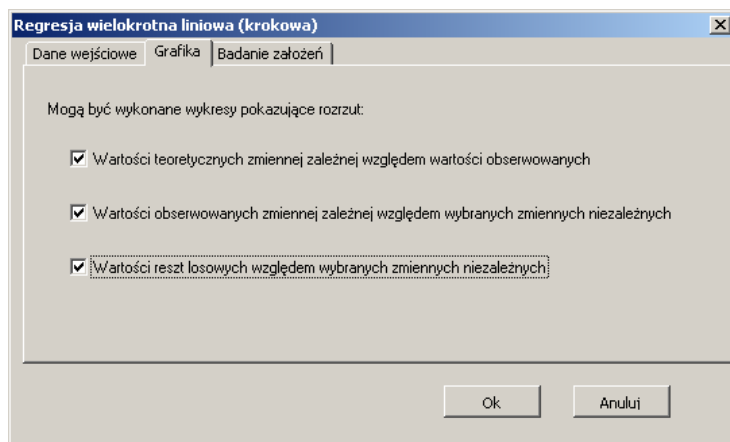
Do estymacji parametrów tego modelu, badania jego istotności, zbadania odpowiednich założeń wymaganych w klasycznej regresji wielokrotnej liniowej wykorzystamy procedurę *Regresja wielokrotna* z menu *Regresja* skoroszytu *StatystykaJG*.



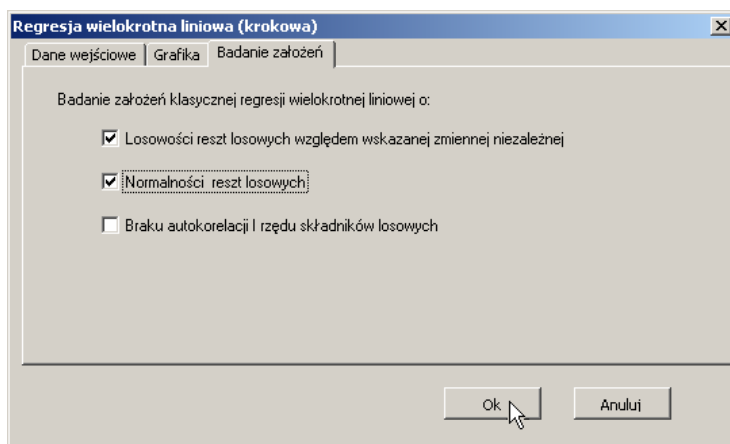
Po wywołaniu procedury *Regresja wielokrotna* zostanie wyświetlone pokazane niżej, kilku zakładkowe okno dialogowe. W zakładce *Dane wejściowe* wskazujemy obszar danych wejściowych, przy czym pierwszy wiersz wskazywanego obszaru musi zawierać unikalne nazwy zmiennych. Jako miejsce zwrócenia wyników ja wybrałem opcję nowego arkusza. Poziom ufności ustawiony jest na standardowym poziomie 0,95, a dokładność wydruku na 2 miejsca po kropce. Bardzo ważne jest zaznaczenie pola wyboru *Predykcja* z uwagi na zamiar późniejszego prognozowania.



W zakładce *Grafika* możemy zadysponować wykonanie lub możliwość wykonania różnych wykresów.

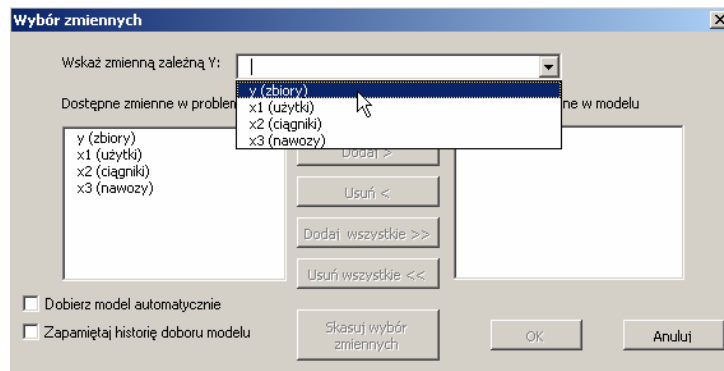


W zakładce *Badanie założeń* zaznaczymy dwa pierwsze pola wyboru (bez badania braku autokorelacji I rzędu).

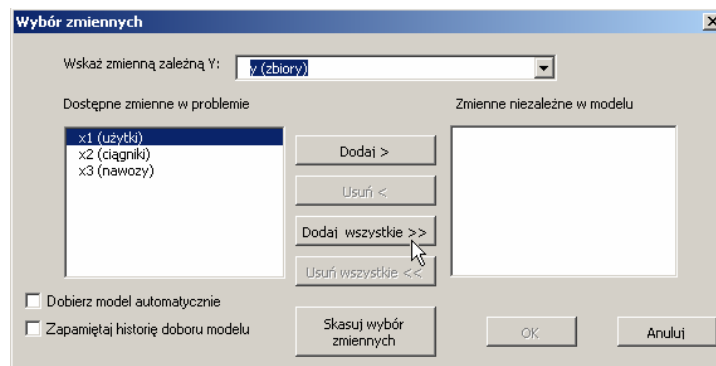


Klik przycisku *Ok* uruchamia proces obliczeniowy, jego pierwszy etap polega na przygotowaniu macierzy sum kwadratów odchyleń i iloczynów odchyleń oraz wektora średnich dla wszystkich zmiennych wskazanych w obszarze danych wejściowych. Po zakończeniu tych prac procedura wyświetla kolejne okno dialogowe w celu ustalenia roli zmiennych w modelu. W oknie tym wskażemy zmienną zależną oraz zmienne niezależne.

Określenie roli zmiennych zaczynamy od wskazania tej zmiennej, która będzie traktowana jako zmienna zależna, zmienna objaśniana. Zmienną tę wybieramy w rozwijanej liście *Wskaż zmienną zależną Y*, w pokazanej sytuacji jest to zmienna  $y$  (zbiory).



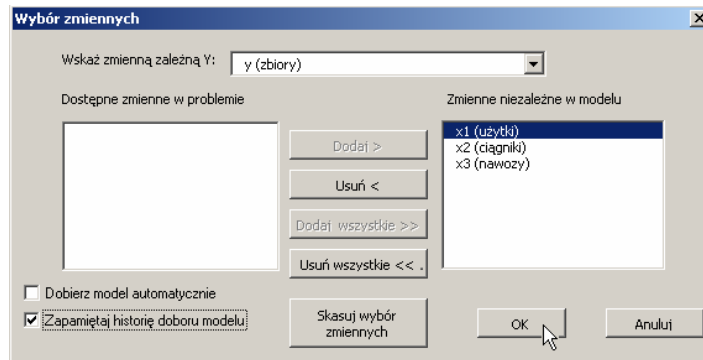
Po wyborze zmiennej zależnej wskazujemy te zmienne spośród dostępnych, które chcemy mieć w modelu jako zmienne niezależne. Można to zrobić przyciskiem *Dodaj >* po wcześniejszym wskazaniu zmiennej w liście *Dostępne zmienne w problemie*, można także skorzystać z przycisku *Dodaj wszystkie >>*, jeżeli wszystkie pozostałe zmienne chcemy mieć w modelu. Dokładnie tak postąpiłem i ja w pokazanej niżej sytuacji.



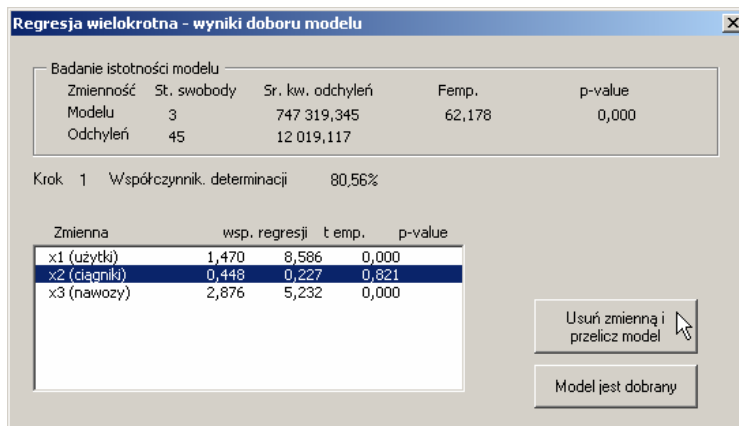
W przypadku ewentualnej pomyłki z wyborem zmiennych możemy skorzystać z dostępnych przycisków korygujących nasz wybór.

Przed uruchomieniem dalszych obliczeń (co będzie możliwe dopiero po wskazaniu zmiennych) zaznaczmy jeszcze pole wyboru *Zapamiętaj historię doboru modelu*, chodzi po prostu o to, aby w momencie zwracania wyników obliczeń procedura zwróciła szczegóły poszczególnych etapów regresji krokowej. Nie polecam zaznaczania pola

wyboru *Dobierz model automatycznie*, nie mamy wtedy kontroli nad przebiegiem regresji krokowej.

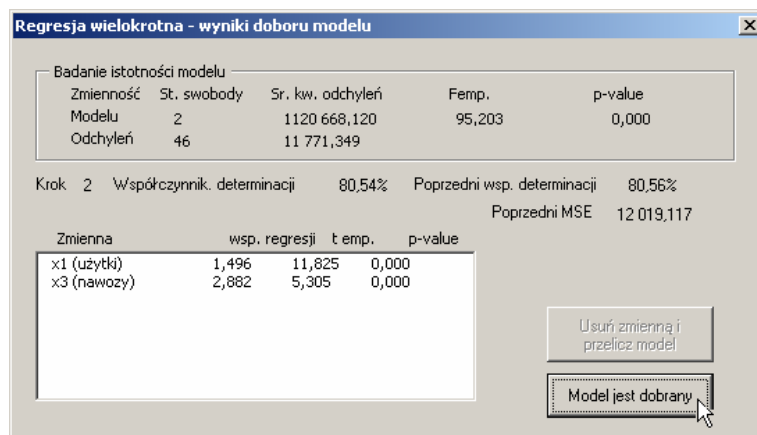


Klik przycisku Ok uruchamia obliczenia regresji krokowej, jej wynik będą zwracane w specjalnym oknie dialogowym. Poniżej pokazane są wyniki pierwszego kroku regresji krokowej, w modu zostały uwzględnione wszystkie trzy zmienne niezależne, model jako taki jest istotny statystycznie, ale w serii weryfikacji trzech hipotez zerowych o istotności poszczególnych częściowych współczynników regresji zmienna x2 (ciągniki) została wskazana jako ta, która może być z modelu usunięta w tym kroku jako zmienna nieistotna. Klik przycisku *Usuń zmienną i przelicz model* spowoduje jej usunięcie i ponowne przeliczenie modelu.



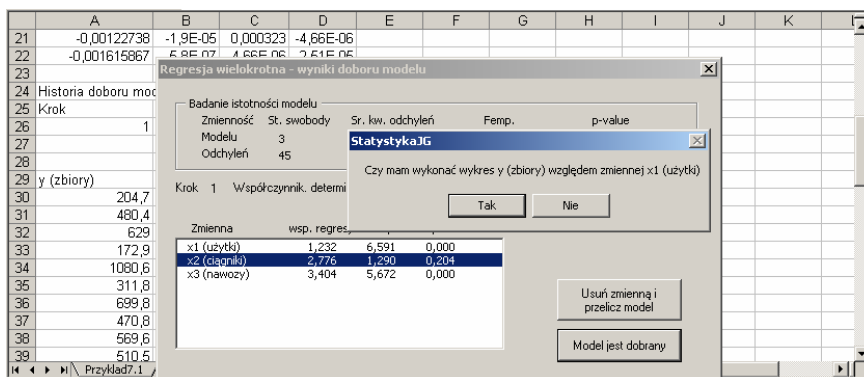
Po usunięciu wskazanej zmiennej ponownie wyświetlane jest okno dialogowe z wynikami regresji krokowej. Zgodnie z oczekiwaniem zmniejszył się współczynnik determinacji, ale zmiana ta jest bardzo mała: z 80,56% na 80,54%, a więc zaledwie o 0,02%. W przypadku średniego kwadratu odchyłeń dla zmienności resztowej też odnotowujemy

jego zmniejszenie: nowy średni kwadrat (wyliczony po usunięciu zmiennej x2) jest mniejszy niż poprzedni – 11771,349 wobec 12019,117.



W drugim kroku regresji krokowej mamy już model dobrany, pozostaje nam jedynie klik przycisku *Model jest dobrany*, co uruchomi procedurę zwrócenia wyników estymacji modelu wraz ze sporządzeniem odpowiednich wykresów (jeżeli taka opcja została zaznaczona) oraz weryfikacją założeń (jeżeli zostały zadysponowane).

W trakcie zwracania wyników obliczeń mogą być wyświetlane komunikaty z pytaniami o wykonanie lub nie danej operacji. W pokazanej niżej sytuacji jest pytanie o wykonanie wykresu rozrzutu punktów zmiennej y (zbiory) względem zmiennej niezależnej x1 (użytki). Pytanie to będzie powtarzane dla wszystkich zmiennych niezależnych.



Podobne pytania mogą być zadawane odnośnie wykonania lub nie wykresów dla reszt względem poszczególnych zmiennych niezależnych.

21	-0,00122738	-1,9E-05	0,000323	-4,66E-06							
22	-0,001615867	-5,8E-07	-4,66E-06	2,51E-05							
23											
24	Historia doboru mo	Regresja wielokrotna - wyniki doboru modelu									
25	Krok	Badanie istotności modelu									
26		Zmiennosc	St. swobody	Sr. kw. odchyleń	Femp.	p-value					
27		Modelu	3				Statystyka JG				
28		Odchyleń	45				Czy mam wykonać wykres reszt względem zmiennej x1 (uzytki)				
29	y (zbiory)	Krok 1 Współczynnik, determina									
30	204,7	Zmienna      wsp. regresji									
31	480,4	x1 (uzytki)	1,232	6,591	0,000						
32	629	x2 (ciągniki)	2,776	1,290	0,204						
33	172,9	x3 (nowozy)	3,404	5,672	0,000						
34	1080,6	Usun zmienną i przelicz model									
35	311,8	Model jest dobrany									
36	699,8										
37	470,8										
38	569,6										
39	510,5										
40	450,6										

W zależności od tego, co wybraliśmy w zakładce *Badanie założeń* w pierwszym oknie dialogowym regresji wielokrotnej mogą być wyświetlane komunikaty z pytaniami, czy przeprowadzić weryfikację hipotezy o poprawności doboru modelu względem poszczególnych zmiennych niezależnych.

21	-0,00122738	-1,9E-05	0,000323	-4,66E-06							
22	-0,001615867	-5,8E-07	-4,66E-06	2,51E-05							
23											
24	Historia doboru mo	Regresja wielokrotna - wyniki doboru modelu									
25	Krok	Badanie istotności modelu									
26		Zmiennosc	St. swobody	Sr. kw. odchyleń	Femp.	p-value					
27		Modelu	3				Statystyka JG				
28		Odchyleń	45				Czy mam zweryfikować hipotezę o poprawności doboru modelu względem zmiennej x1 (uzytki)				
29	y (zbiory)	Krok 1 Współczynnik, determina									
30	204,7	Zmienna      wsp. regresji									
31	480,4	x1 (uzytki)	1,232	6,591	0,000						
32	629	x2 (ciągniki)	2,776	1,290	0,204						
33	172,9	x3 (nowozy)	3,404	5,672	0,000						
34	1080,6	Usun zmienną i przelicz model									
35	311,8	Model jest dobrany									
36	699,8										
37	470,8										
38	569,6										
39	510,5										
40	450,6										

Po zakończeniu udzielania odpowiedzi na pomocnicze pytania do nowego arkusza zostają zwrócone wszystkie zadysponowane wyniki estymacji modelu wraz z odpowiednimi wykresami oraz wynikami badania założeń. Poniżej pokazuję pierwszy fragment okna wyników estymacji.

Po podaniu nazwy zmiennej zależnej wyprowadzone są oceny parametrów modelu; mamy ocenę parametru i błąd tej oceny, dolną i górną granicę przedziału ufności dla prawdziwej wartości cząstkowego współczynnika regresji w populacji oraz wartości statystyk *t*-Studenta i *p*-value dla weryfikacji hipotez zerowych o tym, że dany współczynnik jest równy zero i przy dwustronnej alternatywie.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N				
1	Wyniki estymacji modelu regresji wielokrotnej										Badanie poprawności doboru modelu dla zmienn							
2	Zmienna zależna: y (zbiory)										Badanie losowości reszt							
3	Macierz ocen parametrów, błędy standardowe, dolna i górna granica ufności, t empiryczne, p-value										Liczba ser 27							
4	b(i)	Sb(i)	dgu(i)	ggu(i)	t(i)	p-value(i)								Dodatnich 22				
5	Stała	-263,76	61,36	-387,27	-140,26	-4,30	0,00								Ujemnych 27			
6	x1 (użytki)	1,50	0,13	1,24	1,75	11,83	0,00								Konieczne było przejście na zmienną z			
7	x3 (nawozy)	2,88	0,54	1,79	3,98	5,31	0,00								Zemp = 0,512211			
8											Wniosek: Reszty są losowe, model jest popraw							
9	Badanie istotności regresji testem F Fishera-Snedecora																	
10	Zmienność	St. swobod	Sr. kw. odch.	Femp.	p-value													
11	Modelu	2	1120668,12	95,20	4,46E-17													
12	Resztowa	46	11771,349															
13											Badanie losowości reszt							
14	Wsp. korelacji	0,897																
15	Wsp. determinacji	80,5%																
16											Liczba ser 23							
17	Elementy macierzy odwrotnej V0 niezbędnej do prognozowania										Dodatnich 22							
18	0,319802355	-0,00045	-0,0016336															
19	-0,000451154	1,36E-06	-8,527E-07															
20	-0,001633556	-8,5E-07	2,507E-05															
21											Ujemnych 27							
22	Historia doboru modelu metodą regresji krokowej wstecznej										Konieczne było przejście na zmienną z							
23	Krok	Zmiennych	St. sw. błęd	Sr. kw. błę	Wsp. dete	Zmienna usunięta								Zemp = -0,65515				
24	1	3	45	12019,12	0,805643								Wniosek: Reszty są losowe, model jest popraw					
25	2	2	46	11771,35	0,80542	x2 (ciagniki)								W0: X ma rozkład N(m=0; sigma=108,5)				
26											W emp. = 0,994582							
27											W 0.05 0,947							
28											Wniosek: Brak podstaw do odrzucenia hipotezy							

Jak widzimy ocena cząstkowego współczynnika regresji stojącego przy zmiennej x1 (użytki) jest równa 1,50, a błąd tej oceny jest równy 0,13. Współczynnikiw temu można nadać następującą interpretację: jeżeli wielkość użytków rolnych wzrośnie o 1 jednostkę (czyli 1000 ha), to średnio wielkość zbioru ziów wzrośnie o 1,5 jednostki (czyli 1500 ton), przy warunku, że wielkość nawożenia NPK nie ulegnie zmianie.

Uzyskany przedział ufności można zaś zinterpretować następująco: mamy 95% pewność, że cząstkowy współczynnik regresji zmiennej x1 jest nie mniejszy niż 1,24, ale nie większy niż 1,75.

W przypadku zmiennej x3 (nawozy) ocena cząstkowego współczynnika regresji jest równa 2,88 z błędem  $\pm 0,54$ , co pozwala na następującą interpretację merytoryczną: jeżeli wielkość nawożenia NPK wzrośnie o 1 jednostkę (1 kg NPK/ha), to wielkość zbioru ziów wzrośnie średnio o 2,88 jednostki (czyli 2,88 tys. ton), przy warunku, że wielkość użytków rolnych nie ulegnie zmianie. Przedział ufności dla tego współczynnika interpretujemy analogicznie jak poprzedni: z 95% pewnością mamy prawo oczekiwać, że cząstkowy współczynnik regresji zmiennej x3 jest nie mniejszy niż 1,79, ale nie większy niż 3,98.

Dla wszystkich trzech współczynnikiw mamy wartość statystyki t-Studenta dla weryfikacji hipotez  $H_0 : b_i = 0$  (dla  $i = 0, 1, 2$ ) wobec alternatyw  $H_1 : b_i \neq 0$  wraz z krytycznym poziomem istotności (p-value). W każdym z trzech przypadków ten krytyczny poziom istotności jest mniejszy od przyjętego poziomu istotności  $\alpha = 0,05$ , tym samym hipotezy zerowe odrzucamy na korzyść alternatyw.

Badanie istotności regresji wielokrotnej przeprowadzane jest testem F Fishera, a weryfikowana hipoteza ma postać  $H_0 : \wedge_i b_i = 0$  (dla  $i = 1, 2$ ) wobec  $H_1 : \vee_i b_i \neq 0$ . W naszym przypadku hipotezę zerową odrzucamy na korzyść alternatywy. Łącząc oba

badania (o istotności regresji) i istotności cząstkowych współczynników regresji mamy prawo napisać, że ocena funkcji regresji z próby ma postać:

$$\hat{m}(x_1, x_3) = -263,76 + 1,50x_1 + 2,88x_3 .$$

Badanie założeń klasycznej liniowej regresji wielokrotnej objęło dwa elementy. Po pierwsze dla każdej ze zmiennych niezależnych występujących w modelu testem serii weryfikowano hipotezę o losowości reszt, wyniki tych weryfikacji wyprowadzone od kolumny K potwierdzają, że model został poprawnie skonstruowany. Po drugie dla wektora reszt została zweryfikowana hipoteza o ich normalności, uzyskany wynik nie przeczy założonej hipotezie, tym samym mamy kolejne potwierdzenie poprawności doboru modelu funkcji regresji.

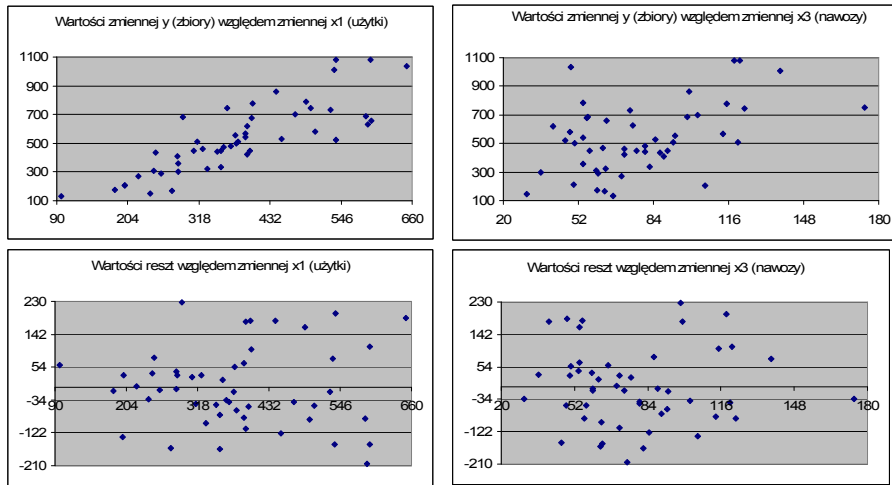
Poniżej wyniki weryfikacji tych trzech hipotez uzupełnione wykresem rozrzutu punktów teoretycznych i obserwowanych wartości zmiennej zależnej.

	B	C	D	E	F	G	H	I	J	K	L	M	N	O														
1	modelu regresji wielokrotnej									Badanie poprawności doboru modelu dla zmiennej x1 (uzytki)																		
2	r (zbiory)									Badanie losowości reszt																		
3	netrów, błędy standardowe, dolna i górna granica ufności, t empiryczne, p-value									Liczbę serii																		
4	b(i)									Dodatnich																		
5	-263,76	61,36	-387,27	-140,26	-4,30	0,00	Ujemnych																					
6	1,50	0,13	1,24	1,75	11,83	0,00	Konieczne było przejście na zmienną z																					
7	2,88	0,54	1,79	3,98	5,31	0,00	Zemp.= 0,512211																					
8										Wniosek: Reszty są losowe, model jest poprawny																		
9																												
10															Badanie poprawności doboru modelu dla zmiennej x3 (nawozy)													
11															Badanie losowości reszt													
12															Liczbę serii													
13															Dodatnich													
14															Ujemnych													
15															Konieczne było przejście na zmienną z													
16															Zemp.= -0,65515													
17															Wniosek: Reszty są losowe, model jest poprawny													
18																												
19																												
20																												
21																												
22	Badanie normalności reszt losowych																											
23	H0: X ma rozkład N(m=0, sigma=108,5)																											
24	W emp.= 0,994582																											
25	W 0.05 0,947																											
26	Wniosek: Brak podstaw do odrzucenia hipotezy zerowej																											
27																												

W przypadku „idealnego” dopasowania punkty te powinny się ułożyć na prostej, oczywiście u nas tak nie ma, ale i wyznaczony współczynnik determinacji z próby równy 80,5% jest znacznie niższy od miary idealnego dopasowania (100%).

Procedura *Regresja wielokrotna* zastosowana do naszego przykładu zwróciła jeszcze 4 dodatkowe wykresy. Pierwsze dwa z nich pokazują rozrzut obserwowanych wartości zmiennej zależnej y-ek od obu zmiennych niezależnych. Wykresy tego typu są pomocne przy konstruowaniu postaci funkcji regresji.

Pozostałe dwa pokazują rozrzut reszt losowych w zależności od obu zmiennych niezależnych. Wykresy te są uzupełnieniem weryfikacji hipotez o losowości reszt względem zmiennych niezależnych.



Na zakończenie tego przykładu zostało nam jeszcze wyznaczenie prognozowanej wielkości zbioru zbóż dla  $x_1=450$ ,  $x_2=50$  i  $x_3=80$ . Jak wiemy wyestymowane równanie regresji nie zawiera zmiennej  $x_2$ , stąd prognoza będzie wykonana bez uwzględniania tej wartości.

Do wykonania prognozy wykorzystam procedurę *Prognoza* z menu *Regresja*, ale przed jej wywołaniem przygotuję w arkuszu 3 (miejscu zwrócenia wyników estymacji modelu) miejsce dla wykonania tej prognozy. W obszarze J2:K3 wpisałem nazwy zmiennych oraz ich wartości w punkcie prognozy, wyniki prognozy będą zwrócone na prawo od tego obszaru.

Poniżej pokazane jest okno tej procedury z wprowadzonymi informacjami.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Wyniki estymacji modelu regresji wielokrotnej										Prognoza	
2	Zmienna zależna: y (zbiory)										x1	x3
3	Macierz ocen parametrów, błędy standardowe, dolna i górna granica ufności, t empiryczne, p-value										450	80
4		b()	Sb()	dgu()	ggu()	t()	p-value()					
5	Stała	-263,76	61,36	-360								
6	x1 (użytki)	1,50	0,13									
7	x3 (nawozy)	2,88	0,54									
8												
9	Badanie istotności regresji testem F Fishera-Sned										Wskaż obszar st. swobody i śr. kw. odchyłeń	
10	Zmienność	St. swobod	Śr. kw. odch.	Femp.						Arkusz3!\$A\$12:\$C\$12		
11	Modelu	2	1120668,12	96						Arkusz3!\$A\$17:\$C\$20		
12	Resztowa	46	11771,349							Arkusz3!\$J\$2:\$K\$3		
13												
14	Wsp. korelacji	0,897										
15	Wsp. determinacji	80,5%										
16												
17	Elementy macierzy odwrotnej V0 niezbędnej do pr											
18		0,319802355	-0,00045	-0,0016336								
19		-0,000451154	1,36E-06	-8,527E-07								
20		-0,001633556	-8,5E-07	2,507E-05								
21												

**Prognozowanie w regresji liniowej i wielokrotnej**

Wskaż obszar oszacowań współczynników regresji: Arkusz3!\$B\$4:\$B\$7

Wskaż obszar st. swobody i śr. kw. odchyłeń: Arkusz3!\$A\$12:\$C\$12

Wskaż obszar macierzy odwrotnej do V0: Arkusz3!\$A\$17:\$C\$20

Wskaż obszar zmiennych niezależnych: Arkusz3!\$J\$2:\$K\$3

Poziom ufności: 95

OK Anuluj

Klik przycisku Ok wyprowadza wyniki prognozy.

	I	J	K	L	M	N	O	P	Q	R	S	
1		Prognoza										
2		x1	x3	Y teor.	Bf. Stand	Dgufn.	Ggufn.	Bf. Pred.	Dgpred.	Ggpred.		
3		450	80	640,11	17,76	604,37	675,86	109,94	418,82	861,41		
4												

Biorąc pod uwagę dwa ostatnie wyniki (Q3 i R3) możemy sformułować następujący wniosek: jeżeli wielkość użytków rolnych będzie ustalona na poziomie 450 jednostek a nawożenie mineralne na poziomie 80 kg, to z 95% pewnością mamy prawo oczekiwać, że wielkość zbioru zbóż będzie nie mniejsza niż 418,82 jednostki, a nie większa niż 861,41 jednostki.