PROSTO O DOPASOWANIU PROSTYCH, CZYLI ANALIZA REGRESJI LINIOWEJ W PRAKTYCE

Janusz Wątroba, StatSoft Polska Sp. z o.o.

W niemal wszystkich dziedzinach badań empirycznych mamy do czynienia ze złożonością zjawisk i procesów. W związku z tym występuje naturalne zainteresowanie ze strony badaczy metodami analizy danych, które umożliwiają ilościową ocenę związków występujących pomiędzy różnymi aspektami badanych zjawisk i procesów. Stosunkowo najczęściej do tego celu wykorzystywane są metody regresji liniowej prostej i wielorakiej. W opracowaniu przedstawiono krótko ideę metody regresji liniowej, sposób jej doboru oraz zagadnienie interpretacji oszacowanego modelu.

W drugiej części zaprezentowano przykłady analiz przeprowadzonych z użyciem narzędzi dostępnych w środowisku programu *STATISTICA*.

Wprowadzenie

Jednym z najczęstszych powodów stosowania metod statystycznej analizy danych jest poszukiwanie przyczyn mających wpływ na interesujące badacza zjawiska. Przykładowo dla ekonomisty może być interesujące stwierdzenie, jakie czynniki kształtują sprzedaż wybranych produktów lub usług. Lekarz jest zainteresowany poszukiwaniem czynników wpływających na stan kliniczny pacjentów, u których zdiagnozowano pewną jednostkę chorobową. W badaniach pedagogicznych celem może być poszukiwanie czynników, które wpływają na wynik egzaminu. Z kolei socjologa może interesować, jakie czynniki są odpowiedzialne za poparcie kandydatów w wyborach. Praktycznie w każdej dziedzinie badań empirycznych można bez trudu podać dalsze przykłady zagadnień stawianych w podobny sposób.

Zazwyczaj mamy do czynienia z sytuacją, w której interesujące nas aspekty badanych zjawisk zależą od całego szeregu czynników, traktowanych jako potencjalne przyczyny (wybór takich potencjalnych przyczyn jest oczywiście łatwiejszy w tych dziedzinach badań, w których istnieje dobrze ugruntowana teoria). Bardzo często trudno jest stwierdzić, w jaki sposób określone przyczyny kształtują wybrane przez badacza lub analityka skutki. Kolejnym problemem jest fakt, iż brane pod uwagę czynniki nie są od siebie niezależne, lecz są nawzajem w różny sposób od siebie uzależnione. W związku z tym badacz świadomie wybiera podejście polegające na uproszczeniu badanych powiązań.

Opisywaną sytuację można przedstawić ogólnie za pomocą zapisu:

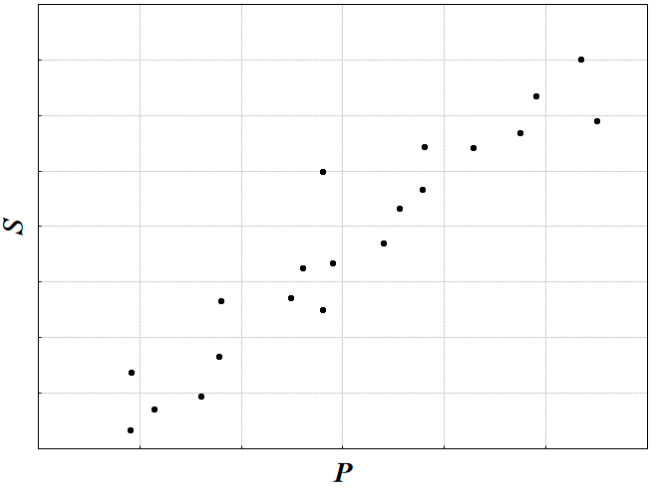
Skutek ← Przyczyna(y)

Bardziej formalny sposób podejścia do rozważanego problemu prowadzi do sformułowania jednokierunkowej zależności w postaci funkcji:

$$S = f(P)$$

Najprostszą postacią takiego równania jest funkcja liniowa, w przypadku której przyjmujemy, że *S* jest proporcjonalne do *P*. Przyjęcie liniowej postaci zależności pozwala w łatwy sposób przedstawić graficznie rozważany problem.

Poniżej na dwuwymiarowym wykresie rozrzutu zaprezentowano przykładowy obraz zależności między wielkościami *S* i *P*. Każdy punkt wykresu oznacza pojedynczy obiekt (obserwację, pomiar).



Rys. 1. Wykres ilustrujący powiązanie pomiędzy wielkościami *S* i *P*.

Położenie punktów na wykresie wskazuje na występowanie wyraźnej prawidłowości (tendencji). Jednocześnie widać, że prawidłowość ta nie może być opisana wyłącznie za pomocą zwykłej funkcji liniowej.

Model regresji liniowej prostej

Jedno z możliwych rozwiązań wskazanego powyżej problemu polega na wprowadzeniu do deterministycznego równania $S = f(P)$ zmiennej losowej ξ i założeniu, że rzeczywista zależność *S* od *P* ma charakter stochastyczny [6]. Zmienna losowa ξ to tzw. składnik losowy, którego zadaniem jest odzwierciedlenie w modelu nieprzewidywanego elementu losowości (związanego np. z ludzkimi zachowaniami), wpływu wielu pominiętych

w modelu zmiennych oraz błędów pomiaru wielkości *S*. W ten sposób otrzymujemy równanie (model), które możemy w ogólnej postaci zapisać jako:

$$Y = f(X, \xi)$$

Jest to ***model regresji liniowej prostej***. W modelu tym *Y* oznacza zmienną *zależną*[1] lub *objaśnianą*, *X* to zmienna *niezależna* lub *objaśniająca*. W klasycznej analizie regresji wprowadza się kilka założeń [6]. Najważniejsze z nich to:

- model zakłada stabilność relacji *f* między badanymi zjawiskami,
- model jest liniowy względem parametrów

  $$Y = \beta_0 + \beta_1 \cdot X + \xi,$$

  gdzie β_0 i β_1 to tzw. parametry strukturalne modelu,
- składnik losowy jest zmienną losową o rozkładzie normalnym $N(0, \sigma^2)$.

Założenie stabilności relacji jest bardzo naturalne. Uchylenie tego założenia prowadzi do modeli o parametrach zmiennych w czasie lub modeli przełącznikowych. Liniowa postać badanej funkcji umożliwia wykorzystanie stosunkowo prostych metod estymacji. Założenie normalności rozkładu składnika losowego pozwala przeprowadzić wnioskowanie statystyczne, ponieważ odpowiednie statystyki mają wówczas pożądane rozkłady (np. *t-Studenta*, *F*).

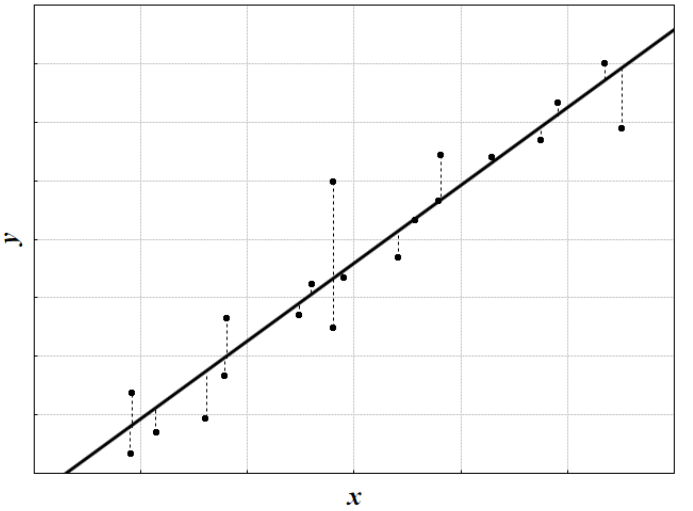
Innymi słowy można powiedzieć, że ze względu na złożoność badanych zjawisk pojawiają się trudności w odwzorowaniu rzeczywistych mechanizmów odpowiedzialnych za ich przebieg. Potrzebne jest zatem uproszczenie. Uproszczone odwzorowanie rzeczywistych współzależności pomiędzy badanymi zjawiskami wymaga od badacza umiejętnego wydobycia istoty mechanizmu generującego dane i przekształcenie go do postaci umożliwiającej zastosowanie ***podejścia statystycznego***. Sprowadza się to do *przyjęcia określonej matematycznej formuły, ujmującej powiązania pomiędzy zmiennymi* oraz założeń o *losowych procesach, wpływających na wyniki pojedynczych pomiarów* [3]. Warto jeszcze raz zwrócić uwagę na to, że przy próbie ilościowego opisu powiązań potrzebne jest rozróżnienie dwóch typów zależności: *deterministycznej* (*funkcyjnej*), która każdej wartości zmiennej *x* przyporządkowuje jedną i tylko jedną wartość zmiennej *y*, oraz *statystycznej* (*probabilistycznej*), która nie przyporządkowuje jednoznacznie wartości *y* danym wartościom *x*, ale może być precyzyjnie opisana za pomocą metod probabilistycznych [4].

Jak dobierana jest linia regresji?

Biorąc pod uwagę rozmieszczeniu punktów na wykresie pokazane na rys. 1, można zaproponować wiele różnych sposobów doboru prostej, która opisywałaby obserwowaną prawidłowość. Najprostsza z tych metod mogłaby polegać na posłużeniu się zwykłą linijką

[1] W książce Maddali [4] na str 96 zamieszczono zestawienie innych nazw używanych dla zmiennych *Y* i *X*.

i dopasowaniu prostej „na oko" w taki sposób, aby poszczególne obserwacje leżały blisko niej. Oczywiście potrzebne jest bardziej formalne kryterium, ale sama idea dopasowania jest właściwie bardzo podobna. Linia regresji będąca graficznym odpowiednikiem modelu regresji jest tak dobierana, aby wielkość będąca sumą kwadratów odległości wszystkich punktów empirycznych od odpowiednich punktów na linii regresji była jak najmniejsza (rys. 2).



Rys. 2. Wykres ilustrujący kryterium doboru linii regresji.

Opisane kryterium jest określane nazwą: ***metoda najmniejszych kwadratów (MNK)***. Kryterium to można formalnie zapisać jako:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \min$$

Praktycznym efektem zastosowania tego kryterium jest możliwość oszacowania parametrów strukturalnych modelu regresji (β_0 i β_1), które charakteryzują się pożądanymi własnościami.

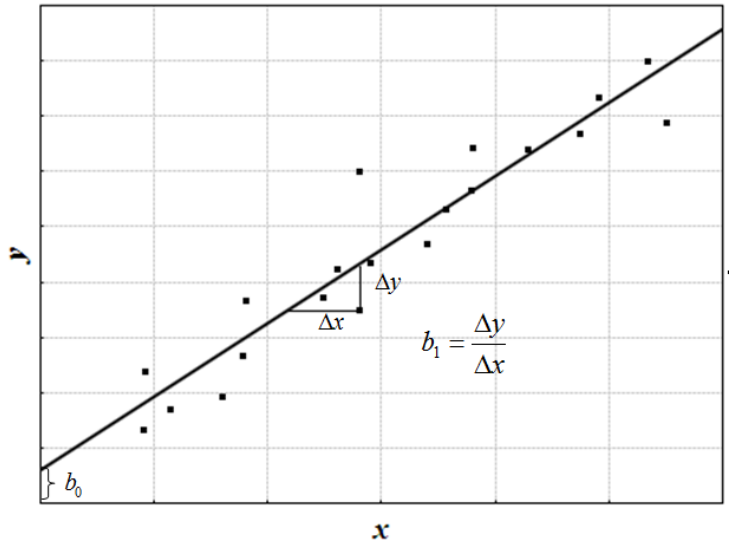
Od czego zacząć interpretację?

Po oszacowaniu parametrów strukturalnych otrzymuje się ich oceny w próbie i w związku z tym model regresji możemy zapisać w postaci:

$$\hat{y} = b_0 + b_1 \cdot x,$$

gdzie $\hat{y}$ oznacza wartość przewidywaną zmiennej zależnej, a b_0 i b_1 to oceny parametrów strukturalnych modelu.

Wielkość b_0 oznacza współrzędną y-ową punktu przecięcia dopasowanej linii regresji z osią OY, natomiast b_1 jest współczynnikiem nachylenia linii regresji do osi OX. Pokazano to na poniższym rysunku.



Rys. 3. Interpretacja ocen parametrów strukturalnych modelu regresji liniowej.

Przy wnioskowaniu statystycznym o parametrach strukturalnych modelu sprawdza się, czy parametry te istotnie różnią się od zera. W tym celu korzysta się z rozkładu statystyki *t-Studenta*. W praktyce większe znaczenie ma ocena istotności parametru β_1, którego ocena z próby mówi o tym, jakiego przeciętnego przyrostu wartości zmiennej zależnej możemy się spodziewać, przy założeniu przyrostu wartości zmiennej niezależnej o 1 jednostkę.

Jak sprawdzić, czy model dobrze pasuje do danych?

Do oceny dopasowania modelu do danych empirycznych stosowanych jest wiele różnych statystyk diagnostycznych. Jedną z najczęściej stosowanych jest **współczynnik determinacji**, oznaczany przez R^2. Oblicza się go ze wzoru:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

gdzie $\hat{y}$ oznacza wartość przewidywaną zmiennej zależnej, a $\bar{y}$ średnią wartość zmiennej zależnej *y*.

Licznik powyższego ułamka określa zmienność wielkości $\hat{y}_i$, a mianownik mierzy zmienność obserwowanych wartości y_i. Współczynnik $\boldsymbol{R^2}$ jest więc miarą stopnia, w jakim model wyjaśnia kształtowanie się zmiennej y. Przyjmuje on wartości z przedziału [0; 1]. Im jego wartość jest bliższa 1, tym dopasowanie modelu do danych jest lepsze.

Inna miara zgodności modelu z danymi empirycznymi opiera się na wariancji składnika losowego. Punktem wyjścia są w tym przypadku tzw. **reszty modelu**. Reszta, która odpowiada i-tej obserwacji, wyraża się wzorem:

$$e_i = y_i - \hat{y}_i, \text{ gdzie i=1, 2, ..., n}$$

Ocena wariancji składnika losowego, tzw. wariancja resztowa, jest obliczana według wzoru:

$$S_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1}$$

gdzie: *n* oznacza liczbę obserwacji, a *k* liczbę zmiennych objaśniających w modelu.

Pierwiastek z wariancji resztowej, czyli odchylenie standardowe reszt $\boldsymbol{S_e}$ (zwany także **błędem standardowym estymacji**), jest powszechnie stosowaną miarą zgodności modelu z danymi empirycznymi. Wielkość ta wskazuje na przeciętną różnicę między zaobserwowanymi wartościami zmiennej objaśnianej i wartościami teoretycznymi. Jest to wielkość mianowana (miano tej wielkości jest takie samo jak zmiennej objaśnianej). Na jej podstawie można również obliczyć miarę niemianowaną, a mianowicie tzw. **współczynnik zmienności losowej**, który określa wzór:

$$W = \frac{S_e}{\bar{y}}$$

Współczynnik ten informuje o tym, jaką część średniej wartości zmiennej objaśnianej stanowi błąd standardowy estymacji, i jest zazwyczaj wyrażany w procentach.

A co z założeniami?

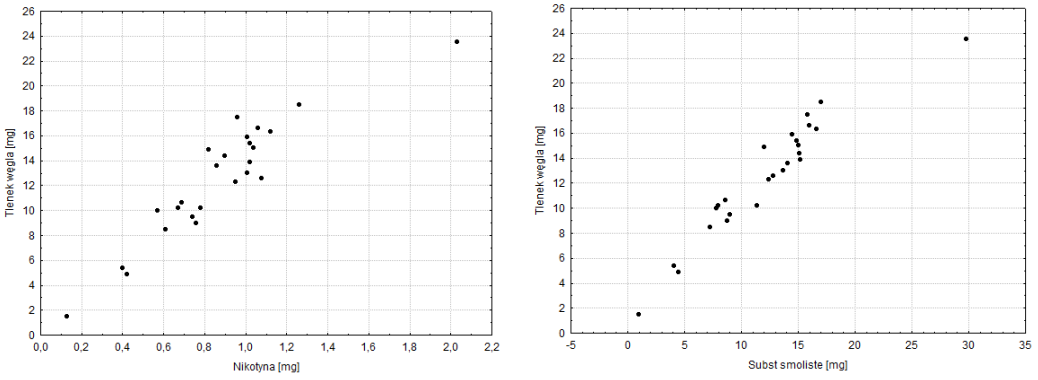
Poprawność wyników analizy regresji zależy od tego, w jakim stopniu są spełnione jej najważniejsze założenia. Wyczerpujący opis oraz dyskusję założeń klasycznej analizy regresji, konsekwencje ich niespełnienia oraz omówienie zalecanych sposobów postępowania można znaleźć w podręczniku Welfego [6]. W niniejszym opracowaniu zwrócimy uwagę na założenia dotyczące składnika losowego (ξ). Najważniejsze z nich dotyczy normalności rozkładu. Jak to zostało już wspomniane wcześniej, spełnienie tego założenia pozwala przeprowadzić wnioskowanie statystyczne, ponieważ odpowiednie statystyki mają wówczas pożądane rozkłady (np. *t-Studenta*, *F*). W części zawierającej opis przykładów analizy regresji zostanie przedstawiony sposób sprawdzania normalności rozkładu składnika losowego.

Przykład analizy regresji liniowej prostej w *STATISTICA*

Dla zilustrowania kolejnych etapów budowy modelu regresji liniowej prostej w środowisku programu *STATISTICA* wykorzystano wyniki oceny 25 marek papierosów różnych producentów, przeprowadzanej corocznie przez Federalną Komisję Handlu w USA [5]. Ocenie podlegały m.in. takie informacje, jak ilość tlenku węgla zawartego w dymie papierosowym oraz zawartość nikotyny i substancji smolistych. Znana jest powszechnie szkodliwość tych substancji dla zdrowia palaczy. Ponadto wyniki badań wskazują na to, że zwiększenie zawartości nikotyny i substancji smolistych wiąże się ze zwiększeniem ilości tlenku węgla w dymie papierosowym.

Dane te posłużyły do wstępnej oceny powiązań występujących pomiędzy branymi pod uwagę zmiennymi oraz budowy modelu regresji liniowej prostej. Ilość tlenku węgla w dymie papierosowym została potraktowana jako zmienna zależna (objaśniana), natomiast zawartość nikotyny i substancji smolistych jako potencjalne zmienne niezależne (objaśniające). Przy okazji został pokazany wpływ jednej nietypowej obserwacji oraz zjawisko współliniowości zmiennych niezależnych.

Przy wstępnej ocenie charakteru i siły badanych powiązań warto posłużyć się dwuwymiarowymi wykresami rozrzutu. Zgodnie z powszechnie przyjmowaną konwencją na wykresie takim na osi *OY* umieszczane są wartości zmiennej zależnej, a na osi *OX* wartości zmiennej niezależnej. Wykresy zostały przedstawione poniżej.



Rys. 4. Powiązanie zawartości tlenku węgla z zawartością nikotyny i substancji smolistych.

Położenie punktów na wykresach wskazuje na występowanie wyraźnego powiązania zawartości nikotyny i substancji smolistych z zawartością tlenku węgla w dymie papierosowym. Ponadto charakter powiązania wskazuje na możliwość dopasowania do danych funkcji liniowej. Jednocześnie na obu wykresach łatwo zauważyć jedną obserwację nietypową (odstającą, skrajną, ang. *outlier*) wyraźnie odbiegającą od pozostałych (powrócimy do tej sprawy w dalszej części opracowania). W kolejnym kroku analizy zostaną zbudowane dwa odrębne modele dla każdej ze zmiennych niezależnych.

W trakcie budowy modelu regresji program *STATISTICA* udostępnia również analityczne narzędzia oceny badanych powiązań. Zamieszczona poniżej tabela zawiera współczynniki korelacji pomiędzy branymi pod uwagę zmiennymi.

Zmienna	Korelacje (Papierosy.sta) Nikotyna [mg]	Subst smoliste [mg]	Tlenek węgla [mg]
Nikotyna [mg]	1,000	0,977	0,926
Subst smoliste [mg]	0,977	1,000	0,957
Tlenek węgla [mg]	0,926	0,957	1,000

Rys. 5. Korelacje pomiędzy zmiennymi.

Otrzymane wartości współczynników korelacji liniowej Pearsona potwierdzają występowanie silnych dodatnich korelacji pomiędzy zawartością tlenku węgla a zawartością nikotyny ($r = 0{,}926$) i substancji smolistych ($r = 0{,}957$). Na tej podstawie możemy stwierdzić, że obydwie analizowane zmienne niezależne mogą być brane pod uwagę jako potencjalne predyktory przy modelowaniu badanych powiązań. Wyniki w tabeli wskazują ponadto na występowanie *współliniowości* zmiennych niezależnych. Na ogół jest ono spowodowane tym, że zmienne charakteryzujące badane zjawiska są ze sobą mocno powiązane lub też jest to związane ze specyfiką zbioru danych, wykorzystywanego do estymacji parametrów modelu regresji. Welfe [2009] rozróżnia dwa rodzaje współliniowości: *dokładną* i *przybliżoną*. Jednym z prostych sposobów postępowania z takimi zmiennymi jest usunięcie jednej ze skorelowanych zmiennych. Omówienie różnych podejść stosowanych w przypadku stwierdzenia silnej współliniowości można znaleźć u Welfego [2009] i Maddali [2006]. W opisywanym przykładzie zbudowano i porównano dwa odrębne modele dla każdej ze zmiennych niezależnych.

N=25	Podsumowanie regresji zmiennej zależnej: Tlenek węgla [mg] (Papierosy.sta) R=0,92595 R^2=0,85738 Skoryg. R^2=0,85118 F(1,23)=138,27 p<0,00000 Błąd std. estymacji: 1,8285					
	b*	Bł. std. z b*	b	Bł. std. z b	t(23)	p
W. wolny			1,66467	0,99360	1,67539	0,10740
Nikotyna [mg]	0,92595	0,07875	12,39541	1,05415	11,75865	0,00000

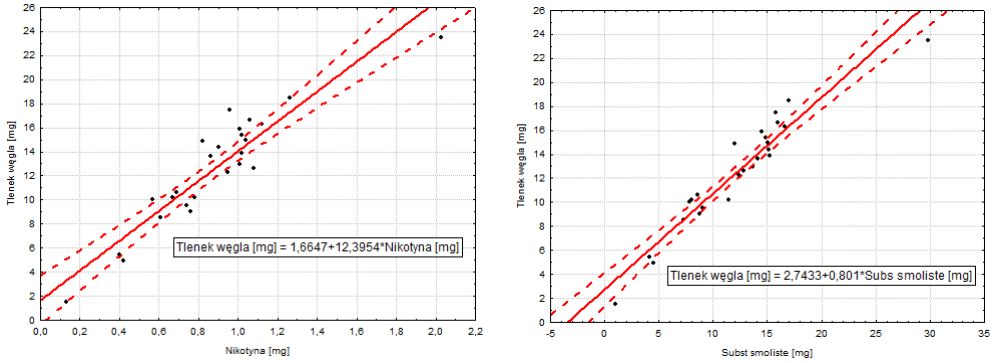
N=25	Podsumowanie regresji zmiennej zależnej: Tlenek węgla [mg] (Papierosy.sta) R=0,95748533 R^2=0,91678 Skoryg. R^2=0,91316 F(1,23)=253,37 p<0,00000 Błąd std. estymacji: 1,3967					
	b*	Bł. std. z b*	b	Bł. std. z b	t(23)	p
W. wolny			2,74328	0,67521	4,06288	0,00048
Subst smoliste [mg]	0,95749	0,06015	0,80098	0,05032	15,91759	0,00000

Rys. 6. Wyniki analizy regresji.

Wyniki analizy pozwalają stwierdzić, że model regresji uwzględniający zmienną niezależną *Nikotyna [mg]* pozwala wyjaśnić ponad 85% wariancji zmiennej *Tlenek węgla [mg]*. Przeciętna różnica pomiędzy rzeczywistymi wartościami zmiennej zależnej i wartościami przewidywanymi przez model wyniosła 1,83 mg (stanowi to 14,6% średniej dla zmiennej zależnej). Wysoka wartość statystyki F (138,27) i odpowiadający jej poziom prawdopodobieństwa p ($p<0{,}001$) potwierdzają statystyczną istotność modelu liniowego. Wartość statystyki t, wykorzystywana do oceny istotności współczynnika regresji (β_1), oraz

odpowiadający jej poziom prawdopodobieństwa *p* potwierdzają, że parametr ten istotnie różni się od zera. Interpretując oszacowaną wartość oceny tego parametru (12,4), możemy stwierdzić, że zwiększenie zawartości nikotyny o 1 mg powoduje zwiększenie zawartości tlenku węgla w dymie papierosowym o 12,4 mg. Z kolei wyraz wolny w modelu (β_0) nieistotnie różni się od zera (oznacza to, że linia regresji przechodzi bardzo blisko środka układu współrzędnych).

Drugi z otrzymanych modeli, uwzględniający zmienną niezależną *Subst smoliste [mg]*, wyjaśnia ponad 91% wariancji zmiennej *Tlenek węgla [mg]*. Tym razem przeciętna różnica pomiędzy rzeczywistymi wartościami zmiennej zależnej i wartościami przewidywanymi była nieco niższa i wyniosła 1,4 mg (stanowi to 11,2% średniej dla zmiennej zależnej). Wysoka wartość statystyki *F* (253,37) i odpowiadający jej poziom prawdopodobieństwa *p* (p<0,001) również potwierdzają statystyczną istotność modelu liniowego. Wartości statystyki *t*, wykorzystywane do oceny istotności współczynnika regresji i wyrazu wolnego, oraz odpowiadające im poziomy prawdopodobieństwa *p* potwierdzają, że parametry te istotnie różnią się od zera. Ponadto otrzymana wartość oceny współczynnika regresji (0,8) pozwala na stwierdzenie, że zwiększenie zawartości substancji smolistych o 1 mg powoduje zwiększenie zawartości tlenku węgla w dymie papierosowym o 0,8 mg. Poniżej zamieszczono również wykresy ilustrujące zbudowane modele.



Rys. 7. Wykresy rozrzutu z dopasowanymi liniami regresji.

Obydwa wykresy potwierdzają bardzo dobre dopasowanie linii regresji (oznaczonych linią ciągłą) do rzeczywistych danych. Ponadto na wykresach zostały również przedstawione krzywe (oznaczone linią przerywaną), wyznaczające 95% przedziały ufności dla wartości oczekiwanych modelowanej zmiennej zależnej.

W trakcie wstępnej analizy danych zauważono wystąpienie jednej obserwacji nietypowej. Zazwyczaj obserwacje takie mają wpływ na wyniki analizy. Poniżej dla porównania zamieszczono tabele z wynikami analizy regresji przeprowadzonej po wykluczeniu nietypowej obserwacji.

N=24	b*	Bł. std. z b*	b	Bł. std. z b	t(22)	p
W. wolny			-0,23818	1,08269	-0,21999	0,82791
Nikotyna [mg]	0,93051	0,07809	14,85998	1,24708	11,91578	0,00000

Podsumowanie regresji zmiennej zależnej: Tlenek węgla [mg] (Papierosy.sta
R=0,93051 R^2=0,86584 Skoryg. R2=0,85974
F(1,22)=141,99 p<0,00000 Błąd std. estymacji: 1,5884
Warunek pomijania: v0='BullDurham'

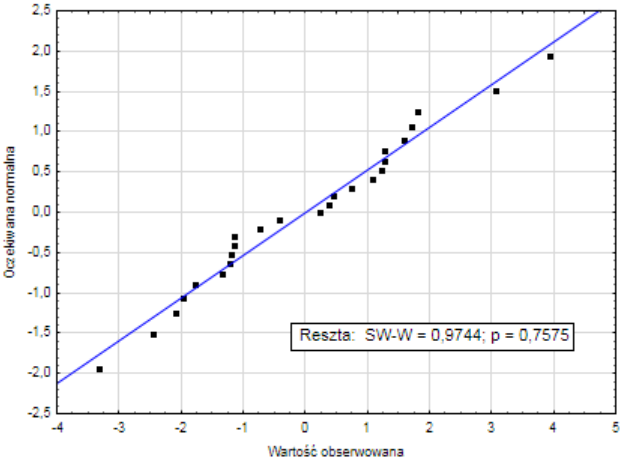
N=24	b*	Bł. std. z b*	b	Bł. std. z b	t(22)	p
W. wolny			1,41285	0,64822	2,17960	0,04027
Subst smoliste [mg]	0,96616	0,05500	0,92813	0,05283	17,56809	0,00000

Podsumowanie regresji zmiennej zależnej: Tlenek węgla [mg] (Papierosy.sta
R=0,96616 R^2=0,93346 Skoryg. R^2=0,93044
F(1,22)=308,64 p<,00000 Błąd std. estymacji: 1,1186
Warunek pomijania: v0='BullDurham'

Rys. 8. Wyniki analizy regresji po usunięciu jednej nietypowej obserwacji.

Otrzymane modele wyjaśniają dodatkowo ponad 1% wariancji modelowanej zmiennej zależnej. Dość znacznym zmianom uległy natomiast oceny wyrazów wolnych i współczynników regresji. Ponadto wyraźnie spadły wartości błędów standardowych estymacji, co oznacza, że modele mają lepsze własności prognostyczne. Należy jednak wyraźnie podkreślić, że usunięcie każdej obserwacji nietypowej musi zawsze być odpowiednio uzasadnione względami merytorycznymi [1].

W ostatniej części przykładu sprawdzimy spełnienie założenia dotyczącego normalności rozkładu składnika losowego. W tym celu utworzono wykres normalności reszt oraz przeprowadzono test Shapiro-Wilka (rys. 9). Wyniki dotyczą modelu uwzględniającego zmienną niezależną *Nikotyna*.



Rys. 9. Wykres normalności reszt i wyniki testu Shapiro-Wilka.

Położenie punktów na wykresie oraz wyniki testu analitycznego wskazują na brak podstaw do kwestionowania normalności rozkładu składnika losowego.

Przykład analizy regresji liniowej wielorakiej

W drugim z prezentowanych przykładów do ilustracji budowy modelu regresji wielorakiej zostanie wykorzystany zbiór danych zawierający wyniki pomiarów procentowej zawartości tkanki tłuszczowej (uzyskane z zastosowaniem techniki ważenia pod wodą) oraz pomiary wybranych cech somatycznych (głównie wymiary obwodów określonych części ciała) zebrane dla 252 dorosłych mężczyzn [2].

Znaczenie zawartości tkanki tłuszczowej w składzie ciała wynika z faktu, iż zbyt wysoka ilość tkanki tłuszczowej może być przyczyną problemów zdrowotnych związanych z układem krążenia, cukrzycą typu II, znacznie podnosi poziom cholesterolu (w konsekwencji prowadzi do miażdżycy) i innych poważnych schorzeń. Natomiast jeżeli poziom tkanki tłuszczowej utrzymywany jest w normie, to człowiek pozostaje w dobrym zdrowiu, ma lepsze samopoczucie, czuje się lekki i szczuplejszy.

Ze względu na trudności z bezpośrednim pomiarem ilości tkanki tłuszczowej opracowano wiele pośrednich metod oceny stanu otłuszczenia. Wszystkie te metody wykorzystują różnego rodzaju pomiary cech budowy ciała lub tworzone na ich podstawie wskaźniki. Merytorycznym celem opisywanego przykładu jest budowa modelu służącego do szacowania procentowej zawartości tkanki tłuszczowej, wykorzystującego pomiary cech budowy ciała otrzymywane z wykorzystaniem prostych narzędzi pomiarowych: wagi i taśmy mierniczej.

Przy budowie modelu regresji pomiar zawartości tkanki tłuszczowej przeprowadzony techniką ważenia pod wodą zostanie potraktowany jako zmienna zależna (objaśniana), a wiek, pomiary wagi i wzrostu oraz obwody jako potencjalne zmienne niezależne (objaśniające).

W przypadku budowania modelu regresji wielorakiej pojawia się problem sposobu doboru i liczby zmiennych objaśniających (niezależnych), które mają zostać uwzględnione w modelu. Liczba zmiennych objaśniających wynika ze znajomości badanej problematyki. Badacz nie powinien tłumaczyć się, że powodem nieuwzględnienia określonej zmiennej objaśniającej była nieznajomość jej wpływu na zmienną objaśnianą (zależną) lub nieodpowiednia wielkość próby czy też niewłaściwy pomiar wartości tej zmiennej. Ważną rzeczą jest skuteczność, a model regresyjny bez zmiennych, które powodują systematyczne zmiany zmiennej zależnej *Y*, jest nieprawdziwy, a ponadto prowadzi do obciążonych estymatorów parametrów modelu. Istotność niektórych zmiennych ustala się metodami statystycznymi, jednak nie można tym zastąpić analizy merytorycznej. Statystyczna analiza zbioru zmiennych objaśniających dotyczy zmniejszania liczby tych zmiennych. Model uwzględniający zbyteczne zmienne charakteryzuje się gorszymi własnościami numerycznymi i jakość estymatorów jest zwykle gorsza z powodu większych błędów i występowania intensywniejszych wzajemnych zależności wśród zmiennych objaśniających.

Wśród metod doboru zmiennych do modelu wyróżniamy: standardową, krokowe, wprowadzania lub usuwania zmiennych oraz wszystkich możliwych regresji. W niniejszym opracowaniu przedstawiono wyniki budowania modelu metodą regresji krokowej wstecznej oraz wszystkich możliwych regresji. W pierwszej z tych metod w pierwszym etapie budowany jest model zawierający wszystkie dostępne zmienne niezależne. Następnie

w kolejnych etapach usuwane są kolejne najmniej istotne zmienne niezależne, aż do uzyskania modelu uwzględniającego tylko zmienne niezależne istotne.

Na samym początku warto przyjrzeć się korelacjom wszystkich zmiennych niezależnych z modelowaną zmienną zależną.

	Korelacje (Otłuszczenie.sta)														
Zmienna	Tłuszcz_% Brozek	Wiek	Waga	Wzrost	BMI	Obwód szyi	Obwód klatki piersiowej	Obwód brzucha	Obwód bioder	Obwód uda	Obwód kolana	Obwód kostki	Obwód bicepsu	Obwód przedramienia	Obwód nadgarstka
Wiek	0,291	1,000	-0,016	-0,246	0,128	0,119	0,181	0,242	-0,058	-0,214	0,017	-0,110	-0,044	-0,085	0,217
Waga	0,620	-0,016	1,000	0,511	0,866	0,806	0,891	0,874	0,929	0,851	0,843	0,581	0,785	0,682	0,719
Wzrost	-0,031	-0,246	0,511	1,000	0,023	0,325	0,223	0,185	0,389	0,343	0,508	0,394	0,318	0,322	0,397
BMI	0,748	0,128	0,866	0,023	1,000	0,744	0,913	0,916	0,859	0,786	0,685	0,453	0,725	0,607	0,604
Obwód szyi	0,483	0,119	0,806	0,325	0,744	1,000	0,766	0,723	0,693	0,656	0,640	0,433	0,707	0,660	0,732
Obwód klatki piersiowej	0,701	0,181	0,891	0,223	0,913	0,766	1,000	0,910	0,821	0,706	0,697	0,448	0,707	0,599	0,641
Obwód brzucha	0,825	0,242	0,874	0,185	0,916	0,723	0,910	1,000	0,859	0,738	0,712	0,408	0,656	0,529	0,595
Obwód bioder	0,638	-0,058	0,929	0,389	0,859	0,693	0,821	0,859	1,000	0,884	0,811	0,519	0,717	0,596	0,609
Obwód uda	0,557	-0,214	0,851	0,343	0,786	0,656	0,706	0,738	0,884	1,000	0,780	0,502	0,740	0,598	0,529
Obwód kolana	0,497	0,017	0,843	0,508	0,685	0,640	0,697	0,712	0,811	0,780	1,000	0,584	0,653	0,575	0,645
Obwód kostki	0,247	-0,110	0,581	0,394	0,453	0,433	0,448	0,408	0,519	0,502	0,584	1,000	0,449	0,429	0,543
Obwód bicepsu	0,482	-0,044	0,785	0,318	0,725	0,707	0,707	0,656	0,717	0,740	0,653	0,449	1,000	0,701	0,611
Obwód przedramienia	0,365	-0,085	0,682	0,322	0,607	0,660	0,599	0,529	0,596	0,598	0,575	0,429	0,701	1,000	0,597
Obwód nadgarstka	0,332	0,217	0,719	0,397	0,604	0,732	0,641	0,595	0,609	0,529	0,645	0,543	0,611	0,597	1,000
Tłuszcz_% Brozek	1,000	0,291	0,620	-0,031	0,748	0,483	0,701	0,825	0,638	0,557	0,497	0,247	0,482	0,365	0,332

Rys. 10. Współczynniki korelacji zmiennej zależnej ze zmiennymi niezależnymi oraz w obrębie zmiennych niezależnych.

Jak widać, stosunkowo najmocniejsze powiązanie z otłuszczeniem ciała wykazuje obwód brzucha (*r=0,825*), BMI (*r=0,748*) oraz obwód klatki piersiowej (*r=0,701*). Jednocześnie widać wyraźnie, że niektóre ze zmiennych niezależnych są również mocno powiązane ze sobą (np. współczynnik korelacji pomiędzy obwodem bioder i wagą wynosi *0,929*). W związku z tym zmienne te będą się nawzajem eliminować w kolejnych etapach budowy modelu.

Poniżej przedstawiono końcowe wyniki ostatecznego modelu, do którego weszły zmienne: *Wiek*, *Obwód brzucha* oraz *Obwód nadgarstka*.

	Podsumowanie regresji zmiennej zależnej: Tłuszcz_% Brozek (Otłuszczenie.sta) R=0,85606 R^2=0,73284 Skoryg. R^2=0,72959 F(3,247)=225,85 p<0,0000 Błąd std. estymacji: 4,0087					
N=251	b*	Bł. std. z b*	b	Bł. std. z b	t(247)	p
W. wolny			-10,59534	5,14068	-2,06108	0,04034
Wiek	0,11717	0,03404	0,07154	0,02079	3,44167	0,00068
Obwód brzucha	0,95183	0,04136	0,71831	0,03122	23,01074	0,00000
Obwód nadgarstka	-0,26062	0,04112	-2,19907	0,34697	-6,33789	0,00000

Rys. 11. Współczynniki korelacji zmiennej zależnej ze zmiennymi niezależnymi oraz w obrębie zmiennych niezależnych.

Na podstawie otrzymanych wyników stwierdzamy, że zbudowany model pozwala wyjaśnić około 73% zmienności modelowanej zmiennej zależnej. Wartość statystyki *F* i odpowiadający jej poziom prawdopodobieństwa testowego *p* potwierdzają istotny statystycznie związek liniowy. Ponadto wartości statystyki *t* wskazują, że wyraz wolny i współczynniki regresji istotnie różnią się od zera.

Interpretując oszacowaną wartość ocen poszczególnych parametrów, możemy stwierdzić, że z każdym rokiem otłuszczenie ciała rośnie przeciętnie o 0,07% (przy niezmienionych wartościach pozostałych zmiennych niezależnych, zasada *ceteris paribus* [1, 4, 6]). Z kolei

zwiększenie obwodu brzucha o jedną jednostkę powoduje zwiększenie otłuszczenia ciała o 0,72% (również przy ustalonych wartościach pozostałych zmiennych). Dość zaskakująco wypada interpretacja oceny współczynnika regresji przy zmiennej *Obwód nadgarstka*. Zwiększenie jej wartości o jedną jednostkę powoduje zmniejszenie otłuszczenia ciała o 2,2% (również przy ustalonych wartościach pozostałych zmiennych).

Przy wykorzystaniu modelu do szacowania rzeczywistego otłuszczenia ciała na podstawie wieku i prostych cech budowy ciała (obwód brzucha i obwód nadgarstka) przeciętny błąd wynosi 4 %.

Pewne ograniczenie podejścia wykorzystującego poszukiwanie metodą regresji krokowej polega na przyjęciu, że istnieje jeden „najlepszy" podzbiór zmiennych niezależnych i poszukiwaniu metody jego identyfikacji. Często zachodzi sytuacja, gdy nie ma jednego „najlepszego" podzbioru. W związku z tym niektórzy statystycy sugerują, że można następnie spróbować dopasować modele metodą wszystkich możliwych regresji, zawierające podobną liczbę zmiennych niezależnych jak w przypadku rozwiązania metodą regresji krokowej, aby zbadać, czy przypadkiem niektóre inne podzbiory zmiennych nie są lepsze. Rozumowanie to sugeruje, że po znalezieniu rozwiązania metodą krokową, powinien zostać zbadany „najlepszy" ze wszystkich możliwych podzbiorów o tej samej liczbie efektów, w celu sprawdzenia, czy rozwiązanie uzyskane metodą krokową jest rzeczywiście „najlepsze".

Poniżej przedstawiono zbiorcze wyniki budowy modeli o liczbie zmiennych niezależnych od 1 do 6. Dla każdej liczby zmiennych niezależnych przedstawiono wyniki trzech najlepszych modeli, przy przyjęciu jako kryterium wartości współczynnika determinacji R^2. Zamieszczona poniżej tabela zawiera informację o wartości współczynnika determinacji dla danego modelu, liczbie uwzględnionych zmiennych niezależnych oraz standaryzowane współczynniki regresji dla zmiennych, które weszły do modelu.

Podsumowanie regresji metodą najlepszego podzbioru dla zmiennej: Tłuszcz_% Brozek (Otłuszczenie.sta)
R kwadrat i standaryzowane współczynniki regresji dla każdego podmodelu

Nr podzb.	R kwadrat	Liczba zmiennych	Wiek	Waga	Wzrost	BMI	Obwód szyi	Obwód klatki piersiowej	Obwód brzucha	Obwód bioder	Obwód uda	Obwód kolana	Obwód kostki	Obwód bicepsu	Obwód przedramienia	Obwód nadgarstka
1	**0,74564**	**6**	**0,0913**		**-0,1047**			**-0,1597**	**1,0593**						**0,0720**	**-0,2181**
2	0,74552	6	0,0928		-0,0993		-0,1101		0,9760						0,0787	-0,1967
3	0,74543	6	0,0907		-0,1054			-0,1629	1,0465					0,0767		-0,2119
4	**0,74302**	**5**	**0,0745**		**-0,1030**			**-0,1309**	**1,0581**							**-0,1898**
5	0,74250	5	0,0751		-0,0989		-0,0808		0,9872							-0,1741
6	0,74227	5	0,0974	-0,3048					1,0719		0,1435					-0,1846
7	**0,74037**	**4**	**0,0819**		**-0,1018**				**0,9498**							**-0,2113**
8	0,73925	4		-0,3797					1,1898					0,0991		-0,1643
9	0,73862	4		-0,3375		0,1463			1,0793							-0,1567
10	**0,73556**	**3**		**-0,2982**					**1,1784**							**-0,1557**
11	0,73494	3			-0,1334				0,9615							-0,1879
12	0,73284	3	0,1172						0,9518							-0,2606
13	**0,72397**	**2**		**-0,4292**					**1,2002**							
14	0,72003	2							0,9722							-0,2473
15	0,71518	2			-0,1893				0,8599							
16	**0,68056**	**1**							**0,8250**							
17	0,56014	1				0,7484										
18	0,49115	1						0,7008								

Rys. 12. Zbiorcze podsumowanie wyników analizy regresji metodą wszystkich możliwych regresji.

Zawarte w tabeli wyniki pozwalają na porównanie różnych modeli o różnej liczbie uwzględnianych zmiennych niezależnych. W ten sposób badacz może na przykład

w stosunkowo łatwy sposób uwzględnić koszty uzyskania danych o poszczególnych zmiennych niezależnych. Jak widać, model zbudowany poprzednio przy pomocy metody krokowej wstecznej znalazł się w tym zestawieniu pod pozycją 12.

Podsumowanie

W rzeczywistych badaniach często podejmowane jest zagadnienie oceny ilościowych związków między różnymi aspektami zjawisk. Celem takich analiz jest zazwyczaj chęć lepszego ich poznania (potwierdzenie lub obalenie formułowanych w teorii hipotez), możliwość przewidywania rozwoju badanych zjawisk lub procesów, czy wreszcie wykorzystanie znajomości ilościowych zależności do symulacji [1]. Dla zrealizowania tak postawionych celów niezbędne jest odwołanie się do teorii badanego zjawiska, dostęp do wyróżnionych w opisie zjawiska danych, znajomość metody umożliwiającej odwzorowanie hipotez teoretycznych za pomocą zgromadzonych danych statystycznych oraz wiedza potrzebna do tego, aby stwierdzić, w jakim stopniu to odwzorowanie się udało.

Literatura

1. Ekonometria i badania operacyjne. Podręcznik dla studiów licencjackich, pod red. naukową M. Gruszczyńskiego, T. Kuszewskiego i M. Podgórskiej (2009), PWN.
2. Johnson R. W. (1996), Fitting Percentage of Body Fat to Simple Body Measurements, Journal of Statistics Education v. 4, n. 1 (www.amstat.org/publications/jse/v4n1/datasets.johnson.html).
3. Krzanowski W. J. (1998), An Introduction to Statistical Modelling, Arnold.
4. Maddala G. S. (2006), Ekonometria, PWN.
5. McIntyre L. (1994), Using Cigarette Data for An Introduction to Multiple Regression, Journal of Statistics Education v. 2, n. 1 (www.amstat.org/publications/jse/v2n1/datasets.mcintyre.html).
6. Welfe A. (2009), Ekonometria. Metody i ich zastosowanie, PWE.

