



## PRZYKŁADY BUDOWY MODELI REGRESYJNYCH I KLASYFIKACYJNYCH

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

Tematyka artykułu obejmuje wprowadzenie do problematyki modelowania statystycznego i jego roli w badaniu mechanizmów różnorodnych zjawisk i procesów, opisywanych przez zgromadzone dane empiryczne. W opracowaniu przedstawiono także przykłady budowania i interpretacji wyników analizy w przypadku modeli regresyjnych i klasyfikacyjnych.

### Wprowadzenie do problematyki modelowania statystycznego

W poznawaniu zjawisk i procesów otaczającej nas rzeczywistości najbardziej powszechnym celem badań jest zazwyczaj odtworzenie mechanizmów odpowiedzialnych za ich przebieg. Wiernie odzwierciedlenie tych mechanizmów nawet w stosunkowo mało skomplikowanych sytuacjach jest zazwyczaj bardzo trudne i wymaga umiejętności wydobywania najbardziej istotnych elementów i zachodzących między nimi powiązań. Takie uproszczone odwzorowanie rzeczywistości (czyli **modelowanie**) jest stosowane w wielu dziedzinach działalności praktycznej i badawczej człowieka. Model jest pojęciem abstrakcyjnym, swoistym pomostem między abstrakcyjnymi sposobami myślenia a realnie istniejącą rzeczywistością. Przedstawia on pewne wyodrębnione, obiektywnie istniejące relacje, które odwzorowuje za pomocą użytecznych reguł, pozwalających upodabniać zachowanie i własności przedstawionego fragmentu rzeczywistości. Dobrze skonstruowany model w adekwatny sposób odtwarza badane obiekty, zjawiska lub procesy i powinien stanowić kompromis między nadmiernym uproszczeniem rzeczywistości a zbyt dużym nagromadzeniem szczegółów [Ostasiewicz 1998].

W zależności od dziedziny badań, przy modelowaniu stosowane są dwa ogólne podejścia: **dedukcyjne** i **indukcyjne**. Pierwsze z wymienionych podejść polega na tym, że w trakcie procesu modelowania wykorzystywane są prawa i fakty znane z teorii. Sytuacja taka występuje najczęściej w tych dziedzinach badań empirycznych, w których jest możliwość oparcia się na solidnych podstawach teoretycznych. W drugim podejściu punktem wyjścia są empiryczne dane pochodzące z eksperymentu lub obserwacji.

Niezależnie od ogólnego podejścia zastosowanego w procesie modelowania zazwyczaj należy brać pod uwagę jeszcze jeden rodzaj uproszczeń. Chodzi o to, że nie zawsze będące



przedmiotem zainteresowania cechy badanych obiektów daje się bezpośrednio zmierzyć. Czasami wartości zmiennych, na podstawie których modeluje się określone zależności stanowią tylko częściowe odwzorowanie natężenia rzeczywistych cech badanych obiektów. Ma to oczywiście wpływ na wiarygodność wniosków, które są wyciągane w oparciu o zbudowany model. Kolejnym źródłem niedokładnego odzwierciedlenia empirycznych danych przez model mogą być błędy pojawiające się w trakcie pomiaru wartości zmiennych.

W praktyce zbudowanie dobrego modelu wymaga od badacza umiejętnego wyodrębnienia istoty mechanizmu, który opisują dane, i przekształcenia go do postaci umożliwiającej zastosowanie podejścia statystycznego. Najczęściej sprowadza się to do przyjęcia określonej matematycznej formuły ujmującej powiązania pomiędzy mierzonymi zmiennymi oraz założeń dotyczących losowych procesów wpływających na wyniki pojedynczych pomiarów. W ten sposób powstaje **statystyczny model** zjawiska. Dopasowanie takiego modelu do określonego zbioru empirycznych daje podstawę do uogólnienia wyników w szerszym kontekście lub do przewidywania wyników w przyszłości, co często prowadzi do lepszego wyjaśnienia badanego zjawiska (Krzanowski 1998).

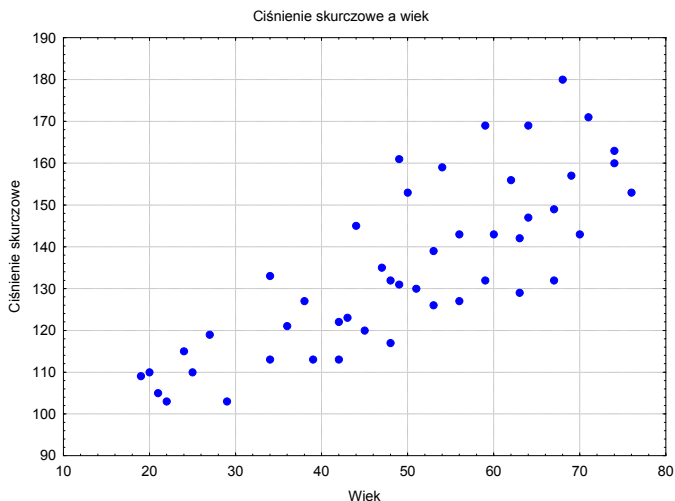
Opisany powyżej bardzo skrótowo sposób formułowania współzależności opisujących interesujące badacza zjawisko lub zespół zjawisk nazywamy **modelowaniem statystycznym**.

W dalszej części opracowania zostaną przedstawione dwa przykłady budowy modeli. Pierwszy z nich będzie dotyczył sytuacji, w której zmienna zależna ma charakter ilościowy. Jest to tzw. zagadnienie regresyjne. Drugi przykład dotyczy budowy modelu dla zmiennej zależnej o charakterze jakościowym. Tego typu zagadnienie nazywane jest zagadnieniem klasyfikacyjnym.

## Przykład budowy modelu regresyjnego w programie *STATISTICA*

Dla zilustrowania procesu budowy modelu dla zmiennej zależnej o charakterze ilościowym wykorzystano dane dotyczące 24 kobiet i 23 mężczyzn w wieku od 19 do 76 lat. Badaną zmienną zależną było skurczowe ciśnienie krwi. Przystępując do analizy zebranych danych, postanowiono ocenić zależność pomiędzy skurczowym ciśnieniem krwi a wiekiem. W tym celu najpierw utworzono wykres rozrzutu. Przedstawia go zamieszczony poniżej rysunek.

Rozmieszczenie punktów na wykresie wskazuje na występowanie pewnej prawidłowości. Widać dość wyraźnie, że wraz z wiekiem stopniowo wzrasta poziom ciśnienia skurczowego krwi. Charakter tej zależności jest zbliżony do przebiegu liniowego. Można również zauważyć zwiększanie się rozrzutu wartości zmiennej zależnej u osób w starszym wieku. Dla ilościowego opisu występującej zależności przeprowadzono analizę regresji liniowej. Najważniejsze wyniki zbudowanego modelu przedstawiono poniżej w tabeli.



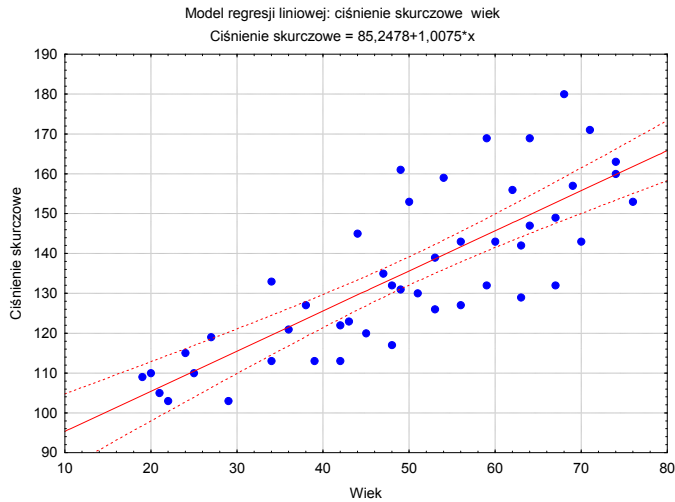
Podsumowanie regresji dla zmiennej zależnej: Ciśnienie skurczowe						
R=0,80739 R2=0,65188 Skoryg. R2=0,64414						
F(1,45)=84,265 p<0,0 Błąd std. estymacji: 12,098						
	BETA	Błąd st. BETA	B	Błąd st. B	t(45)	poziom p
N=47						
W. wolny			85,24783	5,71534	14,91563	0,00000
Wiek	0,80739	0,08795	1,00745	0,10975	9,17962	0,00000

Oszacowany model można zapisać w następującej postaci:

$$\text{Ciśnienie skurczowe} = 85,24783 + 1,00745 * \text{Wiek}$$

Interpretując uzyskaną wartość oceny współczynnika regresji, możemy stwierdzić, że przyrost wieku o jeden rok powoduje wzrost średniego ciśnienia skurczowego krwi o około 1 jednostkę. Dwie najczęściej wykorzystywane statystyki służące do oceny dobroci dopasowanego modelu to współczynnik determinacji  $R^2$  oraz błąd standardowy estymacji  $S_e$ . Współczynnik determinacji mierzy zgodność dopasowania modelu do rzeczywistych danych i informuje, jaka część całkowitej zmienności zmiennej zależnej jest wyjaśniana przez zbudowany model. W naszym przypadku okazuje się, że model pozwala wyjaśnić nieco ponad 65% zmienności ciśnienia skurczowego krwi. Tak więc pozostałe 35% może zostać wyjaśnione przez inne nieuwzględnione w modelu czynniki. Standardowy błąd estymacji wyniósł około 12,1, co oznacza, że gdybyśmy prognozowali wartości zmiennej zależnej w oparciu o oszacowany model, wówczas mylibyśmy się średnio o  $\pm 12,1$  mm Hg. Stanowi to blisko 9% średniej wartości ciśnienia skurczowego krwi. Poniżej zamieszczono wykres ilustrujący dopasowany model.

Przypuszczano, że płeć badanych pacjentów również ma wpływ na poziom ciśnienia skurczowego krwi. W związku z tym w dalszej części analizy zbudowano model uwzględniający zmienną Płeć. Została ona wprowadzona do modelu jako tzw. zmienna zero-jedynkowa. Podobnie jak poprzednio, najważniejsze wyniki dla oszacowanego modelu przedstawiono w tabeli.



N=47	Podsumowanie regresji dla zmiennej zależnej: Ciśnienie skurczowe R=0,89568 R2=0,80224 Skoryg. R2=0,79325 F(2,44)=89,245 p<0,0 Błąd std. estymacji: 9,223					
	BETA	Błąd st. BETA	B	Błąd st. B	t(44)	poziom p
W. wolny			62,36877	5,88432	10,59915	0,000000
Wiek	0,80250	0,06705	1,00135	0,08366	11,96921	0,000000
Płeć	0,38779	0,06705	15,56467	2,69103	5,78391	0,000001

Widać dość znaczną poprawę dopasowania zbudowanego modelu. Pozwala on wyjaśnić nieco ponad 80% oryginalnej zmienności ciśnienia skurczowego. Niecałe 20% to wpływ innych, nieuwzględnionych w modelu zmiennych. Wartość standardowego błędu estymacji zmniejszyła się do około 9,2, co stanowi około 6,8% średniej wartości skurczowego ciśnienia krwi.

Innym sposobem przedstawienia występujących zależności może być zbudowanie dwóch odrębnych modeli dla każdej z grup badanych kobiet i mężczyzn z osobna. Powinno się jednak wcześniej sprawdzić, czy badane grupy pacjentów nie różnią się pomiędzy sobą przeciętnym wiekiem. Aby to ocenić, obliczono odpowiednie średnie wieku (przedstawiono je poniżej w tabeli).

Płeć	Wiek Średnie	Wiek N	Wiek Odch.std
Kobieta	49,33	24	16,279
Mężczyzna	49,74	23	16,589
Ogół grp.	49,53	47	16,253

Jak widać, zróżnicowanie przeciętnego wieku badanych kobiet i mężczyzn okazuje się stosunkowo niewielkie.

W następnym kroku zbudowano dwa odrębne modele: jeden dla badanej grupy kobiet i jeden dla badanej grupy mężczyzn.



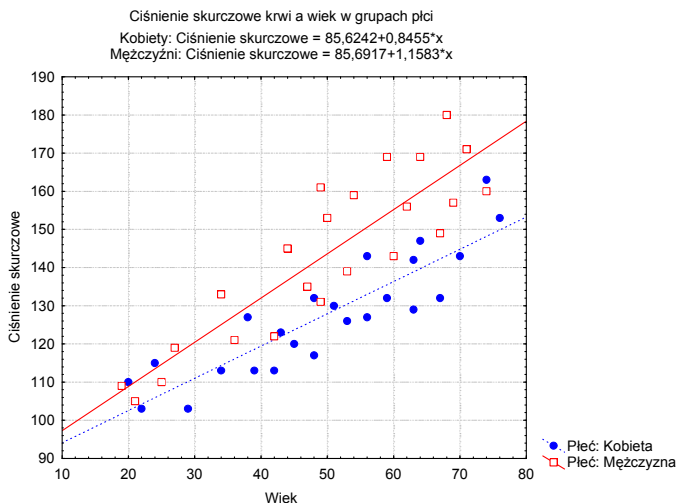
Podsumowanie regresji dla zmiennej zależnej: Ciśnienie skurczowe							Podsumowanie regresji dla zmiennej zależnej: Ciśnienie skurczowe						
R=0,88735 R2=0,78739 Skoryg. R2=0,77773 F(1,22)=81,478 p<0,0 Błąd std. estymacji: 7,313 Włącz warunek: v3=1 (Kobiety)							R=0,88416 R2=0,78174 Skoryg. R2=0,77134 F(1,21)=75,214 p<0,0 Błąd std. estymacji: 10,392 Włącz warunek: v3=2 (Mężczyźni)						
N=24	BETA	Błąd st. BETA	B	Błąd st. B	t(22)	poziom p	N=23	BETA	Błąd st. BETA	B	Błąd st. B	t(21)	poziom p
W. wolny			85,62419	4,85585	17,63320	0,000000	W. wolny			85,69172	6,98754	12,26351	0,000000
Wiek	0,88735	0,09831	0,84546	0,09366	9,02651	0,000000	Wiek	0,88416	0,10195	1,15830	0,13356	8,67261	0,000000

Zbudowane modele można przedstawić w następującej postaci:

$$\text{Kobiety: Ciśnienie skurczowe} = 85,62419 + 0,84546 * \text{Wiek}$$

$$\text{Mężczyźni: Ciśnienie skurczowe} = 85,69172 + 1,15830 * \text{Wiek}$$

Jak widać, zbudowane modele różnią się przede wszystkim wartościami ocen współczynników regresji. Interpretując te wartości, możemy stwierdzić, że w przypadku kobiet przyrost wieku o jeden rok powoduje przeciętny przyrost poziomu skurczowego ciśnienia krwi o około 0,85, podczas gdy w przypadku mężczyzn przyrost wynosi około 1,16. Obydwa oszacowane modele przedstawiono na zaprezentowanym poniżej wykresie.



Jak widać, dopasowane linie regresji różnią się kątem nachylenia. Oznacza to, że w przypadku mężczyzn (u których wartość współczynnika regresji jest wyższa) można zaobserwować większy wzrost przeciętnych wartości skurczowego ciśnienia krwi z wiekiem.

Oszacowane modele można wykorzystać do próby opisanie mechanizmu wpływu wieku na kształtowanie się skurczowego ciśnienia krwi oraz przewidywania jego wielkości przy danym wieku pacjenta, co może mieć duże znaczenie w działaniach profilaktycznych.

## Przykład budowy modelu klasyfikacyjnego w programie *STATISTICA*

Dруга z prezentowanych technik modelowania jest wykorzystywana do ilościowego opisu wpływu zmiennych niezależnych na zmienną zależną, która ma charakter jakościowy. Ma



ona zastosowanie w rozwiązywaniu zagadnień klasyfikacyjnych (zmienna zależna ma charakter zmiennej nominalnej).

W prezentowanym dalej przykładzie wykorzystano dane pochodzące z badań przeprowadzonych wśród mężczyzn w wieku od 15 do 65 lat. Celem tych badań było określenie natężenia wybranych czynników wpływających na występowanie choroby niedokrwiennej serca oraz ocena wpływu tych czynników na poziom ryzyka choroby wieńcowej. Badaniami objęto 210 mężczyzn, u 109 badanych stwierdzono wystąpienie zawału serca, a 101 pozostałych mężczyzn stanowiło grupę porównawczą. Dla każdego z badanych zebrano informacje o wystąpieniu choroby serca oraz o czynnikach, które zwiększają ryzyko jej wystąpienia: występowanie chorób serca w rodzinie badanego, wiek i poziom cholesterolu (LDL).

W pierwszej części analizy oceniono częstość występowania zawału serca w grupach pacjentów, u których występowały bądź nie występowały choroby serca w rodzinie. Do tego celu utworzono tabelę dwudzielczą. Zmieszczono ją poniżej

Choroba wieńcowa	Podsumowująca tabela dwudzielcza: częś		
	Choroby serca w rodzinie Nie	Choroby serca w rodzinie Tak	Wiersz Razem
Tak	43	66	109
%Kolumny	37,39%	69,47%	
Nie	72	29	101
%Kolumny	62,61%	30,53%	
Ogół	115	95	210

Jak widać, u mężczyzn, u których stwierdzano występowanie chorób serca w rodzinie, zawał występował ponad dwa razy częściej.

W dalszej części analizy oceniono zróżnicowanie przeciętnego poziomu pozostałych dwóch zmiennych niezależnych, tzn. wieku i poziomu cholesterolu (frakcja LDL), w grupach pacjentów, u których stwierdzono bądź nie stwierdzono wystąpienia zawału serca. Tabela poniżej przedstawia uzyskane wyniki.

Tabela przekrojów statystyk opisowych (Ryzyko choroby wieńcowej) N=210 (Zmienne zależne nie zawierają BD)						
Choroba wieńcowa	Wiek Średnie	Wiek N	Wiek Odch.std	Poziom cholesterolu (LDL) Średnie	Poziom cholesterolu (LDL) N	Poziom cholesterolu (LDL) Odch.std
Tak	50,82569	109	10,12683	5,502569	109	2,222030
Nie	38,44554	101	14,60512	4,310495	101	1,605757
Ogół grp.	44,87143	210	13,91043	4,929238	210	2,035079

Jak widać, w grupie mężczyzn z zawałem serca można zaobserwować wyższą średnią wieku (o około 12 lat) oraz wyższy przeciętny poziom stężenia cholesterolu (o około 1,2) w stosunku do pacjentów, u których nie wystąpił zawał serca.

Dla oceny łącznego wpływu branych pod uwagę zmiennych niezależnych na wystąpienie zawału serca zastosowano uogólniony model liniowy. Najważniejsze wyniki oszacowanego modelu przedstawiono w poniższej tabeli.

		Choroba wieńcowa - Oceny parametrów (Ryzyko choroby wieńcowej)					
		Rozkład: DWUMIANOWY					
		F. wiążąca: LOGIT					
Efekt	Poziom Efekt	Kolumna	Ocena	Standard Błąd	Wald Stat.	p	Ryzyko względne
<b>W.wolny</b>		1	-4,34956	0,729170	35,58228	0,000000	
Wiek		2	0,06509	0,013433	23,47546	0,000001	1,07
Poziom cholesterolu (LDL)		3	0,20085	0,087236	5,30091	0,021314	1,22
Choroby serca w rodzinie		4	1,09669	0,327825	11,19134	0,000822	2,99
Skala			1,00000	0,000000			

Widać, że wszystkie uwzględnione zmienne niezależne wykazują istotny wpływ na wystąpienie zawału serca. Oceniając wpływ poszczególnych zmiennych, wykorzystuje się tzw. ryzyko względne (Agresti 1996, Lindsey 1995). Wskaźnik ten został podany w ostatniej kolumnie powyższej tabeli. Na tej podstawie możemy stwierdzić, że przyrost wieku o 1 rok powoduje wzrost ryzyka wystąpienia choroby niedokrwiennej serca o około 7% (przy interpretacji bierzemy pod uwagę wartość ryzyka na poziomie 1). W przypadku poziomu cholesterolu przyrost jego stężenia o 1 mmol/l powoduje wzrost ryzyka o około 22%. I wreszcie występowanie chorób serca w rodzinie badanego powoduje prawie 3-krotny wzrost ryzyka wystąpienia choroby niedokrwiennej serca.

Jednym z najczęściej stosowanych kryteriów oceny jakości zbudowanego modelu jest porównanie rzeczywistej klasyfikacji badanych ze względu na wartość zmiennej zależnej z klasyfikacją, którą otrzymuje się w wyniku zastosowania zbudowanego modelu. Odsetek zgodności tych dwóch klasyfikacji przedstawia zamieszczona poniżej tabela.

Klasyfikacja przypadków (Ryzyko choroby wieńcowej)			
Ilor. szans: 5,223214			
Log ilor. szans: 1,653113			
Obserwow.	Przewidywane		Procent poprawne
	Tak	Nie	
Tak	81	28	74,31
Nie	36	65	64,36

Na podstawie tych wyników możemy stwierdzić, że model poprawnie klasyfikuje ponad 74% badanych, u których rzeczywiście wystąpiła choroba serca oraz ponad 64% tych, u których choroba serca w rzeczywistości nie wystąpiła.

## Literatura

1. Agresti A., An Introduction to Categorical Data Analysis, John Wiley & Sons.
2. Krzanowski W. J., 1998, An Introduction To Statistical Modelling, Arnold.
3. Lindsey J. K., 1995, Introductory Statistics. A Modelling Approach, Oxford University Press.
4. Statystyczne metody analizy danych, (1998), red. W. Ostasiewicz, Wydawnictwo Akademii Ekonomicznej we Wrocławiu.
5. Hastie T., Tibshirani R., Friedman J., 2001, The Elements of Statistical Learning, Springer.