ANALIZA REGRESJI

Na poprzednich ćwiczeniach omówiliśmy współczynnik korelacji liniowej Pearsona mierzący siłę i kierunek liniowego związku między dwiema zmiennymi losowymi. Na obecnych ćwiczeniach poświęconych regresji liniowej zajmiemy się modelowaniem związku między: zmienną zależną, oznaczaną przez Y i zmienną niezależną, oznaczaną przez X. Model, który tu będziemy opisywać zakłada, że między X i Y zachodzi liniowy związek. Model regresji liniowej opisujący zależność zmiennej Y od X przyjmuje w takiej sytuacji postaci:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

gdzie odpowiednio:

β_0, β_1 - parametry liniowej funkcji regresji

ε - składnik losowy

Wyraz wolny β_0 jest punktem przecięcia linii prostej z osią rzędnych(oś Y), a β_1 jest współczynnikiem kierunkowym, czyli miarą nachylenia linii $\beta_0 + \beta_1 x$ (względem osi odciętych). Składnik losowy reprezentuje losowe zakłócenia funkcyjnego powiązania między wartościami zmiennej zależnej a wartościami zmiennej niezależnej. Składnik ten wyraża wpływ wszystkich czynników, które obok X wpływać mogą na zmienną objaśnianą Y, oraz związany jest z brakiem pełnego dopasowania analitycznej postaci funkcji regresji do rzeczywistego powiązania między analizowanymi zmiennymi. Składnik ten jest losową zmienną, która pozwala na obliczenie dokładności szacunku parametrów liniowej funkcji regresji. Musimy pamiętać, że w rzeczywistości nie są znane parametry β_0, β_1. Możemy je jedynie oszacować na podstawie n-elementowej próby składającej się z par obserwacji (x_i, y_i) dla i = 1, 2,..., n. Oszacowana funkcja regresji przyjmuje wówczas następującą postać:

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

gdzie odpowiednio:

b_0 i b_1 – oceny parametrów β_0, β_1.

e_i - tzw. reszty (zmienna losowa) definiowane jako $e_i = y_i - \hat{y}_i$, czyli różnica między wartością obserwowaną y_i a teoretyczną wyliczoną z modelu $\hat{y}_i$.

Jak jednak znaleźć taką „dobrze dopasowaną" linię prostą? Punktem wyjścia są reszty, a właściwie suma kwadratów reszt opisująca rozbieżność pomiędzy wartościami empirycznymi zmiennej zależnej a jej wartościami teoretycznymi, obliczonymi na podstawie wybranej funkcji. Oszacowania b_0 i b_1 dobieramy tak, aby suma kwadratów reszt osiągnęła minimum. Ta najbardziej znana i stosowana metoda szacowania parametrów linii regresji nosi nazwę **metody najmniejszych kwadratów** (MNK). Nie musimy się martwić o skomplikowane obliczenia występujące w tej metodzie, bowiem wszystkie pakiety statystyczne obliczają oceny współczynników regresji. Tutaj tradycyjnie pokażemy, jak korzystać z pakietu *STATISTICA,* aby uzyskać pełne rozwiązanie problemu regresji. Pakiet ten dysponuje modułem do przeprowadzenia bardziej ciekawych i złożonych analiz. Jest to moduł **Regresja Wielokrotna**. Przy pomocy tego modułu możemy przeprowadzić obliczenia związane z liniową regresją wielokrotną, regresją krokową lub przeprowadzić analizę modeli nieliniowych, które poprzez transformację sprowadzamy do postaci liniowej.

Rozważmy badanie, w którym analizowano powiazanie między obwodem serca a masą ciała dla 15 krów. Jesteśmy zainteresowani równaniem regresji opisującej zależność masy ciała i obwodu serca. Fragment omawianych danych przedstawia poniższa tabelka.

Masa	641	620	633	651	640	666	650	688	680	670	630	665
Obwód	205	212	213	216	217	218	219	221	226	207	222	212

Nas interesują współczynniki modelu Obwód = b_1·Waga + b_0 wyznaczone metodą najmniejszych kwadratów. Dla naszych przykładowych danych otrzymamy następujące arkusze wyników:

Podsumowanie regresji zmiennej zależnej: Obwód serca (Prz
R= ,78912793 R^2= ,62272288 Skoryg. R2= ,59370157
F(1,13)=21,457 p<,00047 Błąd std. estymacji: 4,0362

N=15	b*	Bł. std. z b*	b	Bł. std. z b	t(13)	p
W. wolny			63,38385	32,89123	1,927074	0,076111
Waga	0,789128	0,170356	0,23335	0,05037	4,632217	0,000469
[1]	[2]	[3]	[4]	[5]	[6]	[7]

Stat.podsum.;

statystyka	**Wartość**
R wielorakie	0,78913
Wielorakie R2	0,62272
Skorygowane R2	0,59370
F(1,13)	21,45743
p	0,00047
Błąd std. estymac	4,03617

Rys 1. Arkusz wyników.

Arkusze te pokazują sumaryczne wyniki analizy regresji oraz dodatkowe statystyki. Współczynniki regresji to kolumna oznaczona przez [4]. Pierwszy wiersz to wartość stała b_0, a drugi to współczynnik b_1. Tak więc poszukiwany model ma postać:

$$\text{Obwód} = 0{,}23335 \cdot \text{Waga} + 63{,}38385$$

Jak wiemy w praktyce nie dysponujemy pełną informacją o populacji generalnej. To co mamy, to funkcja regresji wyliczona metodą najmniejszych kwadratów w oparciu o dane z losowej próby. Wiąże się z tym problem oceny rozbieżności między wartościami zmiennej zależnej y_i a wartościami $\hat{y}_i$ wyliczonymi z modelu. Różnice $e_i = y_i - \hat{y}_i$ opisujące tą rozbieżność jak wiemy noszą nazwę reszt. Im reszty będą mniejsze, tym wartości empirycznej y_i będą bliższe wartości $\hat{y}_i$ przewidywanej przez model. To podpowiada, aby jako miarę omawianej rozbieżności potraktować odchylenie standardowe reszt e_i W statystyce bowiem precyzję estymatora mierzy jego wariancja. I tak jest w istocie, wielkość ta zwana błędem standardowym estymacji i oznaczana jako Se informuje o przeciętnej wielkości odchyleń wartości obserwowanych zmiennej zależnej od wartości wyliczonych z modelu (teoretycznych). Odchylenie standardowe reszt mówi nam o stopniu „dopasowania" modelu do danych empirycznych. Im S_e mniejsze tym lepiej dopasowany model. Wartość ta dla naszego modelu jest równa S_e = 4,0362. Oznacza to, że przewidywane wartości zmiennej Obwód różnią się od wartości obserwowanych średnio biorąc o 4,0362.

Możemy więc napisać: $\text{Obwód} = 0{,}23335 \cdot \text{Waga} + 63{,}38385 \pm 4{,}0362$

Wyliczone współczynniki regresji b_0 i b_1 są, jak wiemy, oszacowaniami współczynników regresji dla całej populacji. Nasuwa się więc pytanie, jakim błędem są one obarczone. Odpowiedzi na nie udziela średni błąd szacunku parametru. Jest on oszacowaniem średniej rozbieżności między parametrami modelu a jego możliwymi ocenami. Wartości te są podane w kolumnie oznaczonej przez [5].mamy zatem:

- oceny parametru b_0 odchylają się od tego parametru o S_{b0} = 32,89123
- oceny parametru b_1 odchylają się od tego parametru o S_{b1} = 0,05037.

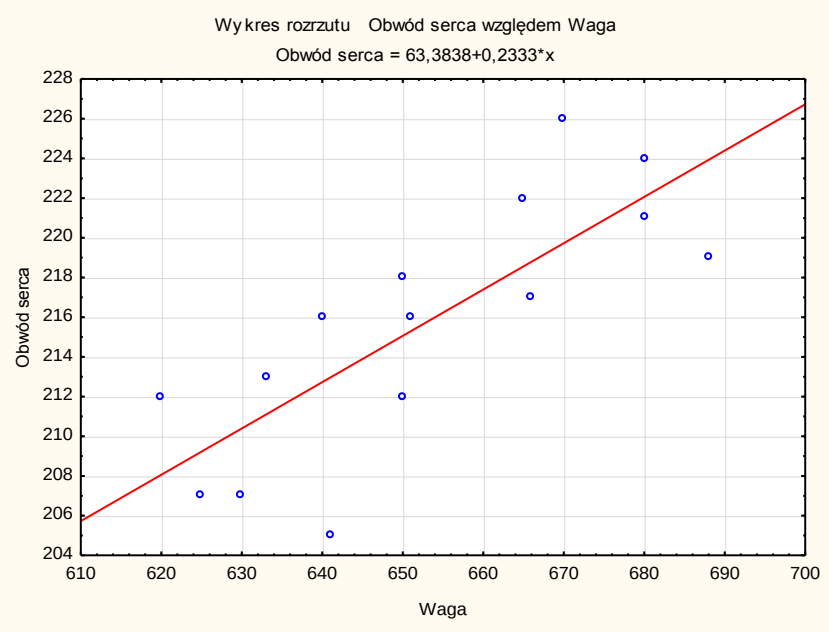
Możemy, więc powiedzieć, że szacując współczynnik kierunkowy na poziomie 0,23335 mylimy się średnio o 0,05037. Podobnie szacując wyraz wolny na poziomie 63,38385 mylimy się średnio biorąc o 32,89123. Przyjęło się zapisywać wielkości S_b w nawiasach pod ocenami parametrów modelu. Dla naszego przykładu mamy zatem:

$$\begin{array}{c} \text{Obwód} = 0{,}23335 \cdot \text{Waga} + 63{,}38385 \pm 4{,}0362 \\ \quad\quad (0{,}05037) \quad\quad\quad (32{,}89123) \end{array}$$

Wielu autorów uważa jednak, że średnie błędy szacunku są niewygodne w użyciu. Dużo łatwiej jest zinterpretować ilorazy t ($t=bi/S_{bi}$) (pole [6]). Weryfikują one istotność analizowanych zmiennych. Jak widzimy w naszym przykładzie tylko parametr dla zmiennej waga jest istotny (pole [7]). Oznacza to, że zmienna Waga ma istotny wpływ na zmienną Obwód.

Omówiliśmy już kilka miar „dopasowania". Najbardziej jednak popularną miarą jest **współczynnik determinacji**. Jest to liczba z przedziału <0, 1>. R^2 równe 1 to doskonałe dopasowanie, natomiast wartość R^2 równe 0 oznacza brak powiązania między zmiennymi. Współczynnik determinacji mierzy nam, jaka część ogólnej zmienności zmiennej zależnej jest wyjaśnione przez regresję liniową. Symbol R^2 wziął się z tego, że w modelu liniowym współczynnik determinacji jest równy kwadratowi współczynnika korelacji. Wartość R^2 znajdziemy w pierwszym i drugim arkuszu wyników (rys. 1). W naszym przykładzie wartość ta wynosi $R^2=0{,}6227$. Można to wyrazić w procentach mówiąc, że model wyjaśnia 62,3% zaobserwowanej zmienności.

Pamiętajmy, im większe R^2 tym lepiej. Nie przesadzajmy jednak. Dołączenie bowiem nowej zmiennej do istniejącego modelu zawsze powoduje zwiększenie R^2. Dlatego w praktyce używamy raczej Skorygowanego R^2. Uwzględnia on, że R^2 jest obliczony z próby i jest trochę „za dobry" jeśli uogólniamy nasze wyniki na populację. Skorygowane R^2 mówi nam jak dobrze dopasowane byłoby nasze równanie regresji do innej próby z tej samej populacji. Poprawione R^2 jest zawsze mniejsze od R^2. Omawiane regresji liniowej kończymy wykresem regresji liniowej dla rozpatrywanego przykładu.



Rys 2. Wykres linii regresji dla analizowanych danych

Rozszerzeniem regresji liniowej jest regresja wieloraka postaci:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \quad + \beta_k X_k + \varepsilon$$

gdzie odpowiednio:

β_0, β_1 …β_k - parametry liniowej funkcji regresji szacowane metoda najmniejszych kwadratów,

ε - składnik losowy.

Jeżeli dodatkowo w naszym przykładzie rozważymy wiek krowy mamy równanie postaci

$$\text{Obwód} = 0{,}183 \cdot \text{Waga} + 2{,}14 \cdot \text{Wiek} + 58{,}346$$