5. Regresja liniowa z pojedynczym i wieloma regresorami

Decyzje zarządcze często opierają się na związku między dwiema lub więcej zmiennymi. Na przykład menedżer ds. marketingu może próbować prognozować sprzedaż przy określonym poziomie wydatków na reklamę po zbadaniu związku między tymi wydatkami a sprzedażą.

W drugim przypadku przedsiębiorstwo publiczne może wykorzystać stosunek między dzienną maksymalną wartością temperatury a zapotrzebowaniem na energię elektryczną, aby przewidzieć zużycie energii elektrycznej. Czasami menedżer polega na intuicji. Intuicyjnie ocenia, w jaki sposób te dwie zmienne są ze sobą powiązane. Jeśli jednak możliwe jest uzyskanie danych, warto zastosować procedurę statystyczną zwaną analizą regresji, aby pokazać, w jaki sposób te dwie zmienne są ze sobą powiązane.

W terminologii regresji zmienna przewidywana nazywana jest zmienną zależną.

Zmienna lub zmienne używane do przewidywania wartości zmiennej zależnej nazywane są zmiennymi niezależnymi.

Analizując wpływ wydatków na reklamę na sprzedaż, sprzedaż byłaby zatem zmienną zależną. Wydatki na reklamę byłyby zmienną niezależną. W notacji statystycznej y oznacza zmienną zależną, a x oznacza zmienną niezależną.

W tej sekcji przyjrzymy się najprostszemu typowi analizy regresji, który obejmuje jedną zmienną niezależną i jedną zmienną zależną. Związek między tymi dwiema zmiennymi będzie przybliżony linią prostą. Nazywa się to prostą regresją liniową. Analiza regresji obejmująca dwie lub więcej zmiennych niezależnych nazywana jest analizą regresji wielorakiej.

5.1 Model prostej regresji liniowej

Best Burger to sieć restauracji szybkiej obsługi zlokalizowanych na obszarze wielu stanów. Lokalizacje Best Burger znajdują się w pobliżu kampusów uniwersyteckich. Menedżerowie uważają, że kwartalna sprzedaż tych restauracji (oznaczona przez y) jest dodatnio skorelowana z wielkością populacji studentów (oznaczonej przez x). Restauracje w pobliżu kampusów z dużą liczbą

studentów mają tendencję do generowania większej sprzedaży niż te w pobliżu kampusów z małą liczbą studentów. Korzystając z analizy regresji, możemy opracować równanie, które pokazuje, w jaki sposób zmienna zależna y jest powiązana ze zmienną niezależną x.

5.2 Model regresji i równanie regresji

W przypadku Best Burger populację stanowią wszystkie restauracje Best Burger. Dla każdej restauracji w populacji istnieje wartość x (populacja studentów) i odpowiadająca jej wartość y (kwartalna sprzedaż). Równanie opisujące, w jaki sposób y jest powiązane z x, nazywane jest **modelem regresji**.

$$y = \beta_0 + \beta_1 x + \epsilon$$

β_0 i β_1 są nazywane parametrami modelu, ϵ (grecka litera epsilon) jest zmienną losową nazywaną błędem modelu. Błąd reprezentuje zmienność y, której nie można wyjaśnić liniową zależnością między x oraz y.

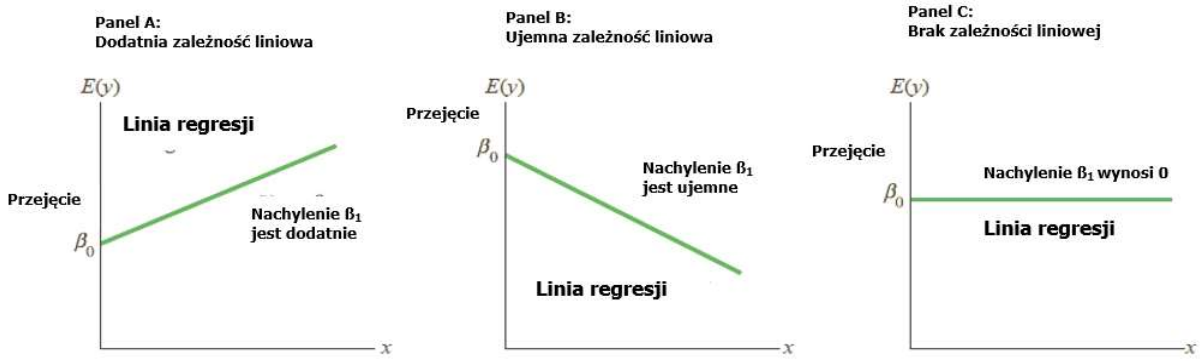
Populacja wszystkich restauracji Best Burger może być również postrzegana jako zbiór subpopulacji, po jednej dla każdej oddzielnej wartości x. Na przykład jedna subpopulacja składa się ze wszystkich restauracji Best Burger w pobliżu kampusów uniwersyteckich z 8000 studentów. Druga subpopulacja składa się ze wszystkich restauracji Best Burger zlokalizowanych w pobliżu kampusów uniwersyteckich z 9000 studentów i tak dalej. Każda subpopulacja ma odpowiadający jej rozkład wartości y. Każdy rozkład wartości y ma swoją średnią lub wartość oczekiwaną. Równanie opisujące wartość oczekiwaną y, oznaczaną jako E(y), która jest powiązana z x, nazywane jest równaniem regresji. Równanie prostej regresji liniowej jest następujące:

$$E(y) = \beta_0 + \beta_1 x$$

Wykresem równania prostej regresji liniowej jest linia prosta. β_0 reprezentuje wartość początkową linii regresji, β_1 to współczynnik kierunkowy linii, a $E(y)$ to wartość średnia lub wartość oczekiwana y dla danej wartości x.

Przykłady możliwych linii regresji pokazano na poniższym rysunku 5.1. Linia regresji w przypadku A pokazuje, że wartość y jest dodatnio skorelowana z x. Wraz ze wzrostem wartości x rosną również wartości $E(y)$. Linia regresji na rysunku B pokazuje wartość y, która jest ujemnie skorelowana z x. Mniejsze wartości $E(y)$ są powiązane z wyższymi wartościami x. Linia regresji na rysunku C pokazuje przypadek, w którym wartość y nie jest powiązana

z x. Oznacza to, że oczekiwana wartość y jest taka sama dla każdej wartości x.



Rysunek 5.1 Przykłady wykresów zależności liniowej

5.3 Szacowane równanie regresji

Gdyby wartości parametrów populacji były znane β_0 i β_1, moglibyśmy użyć powyższego równania do obliczenia wartości y dla danej wartości x. W praktyce parametry te są trudno dostępne, więc są po prostu szacowane na podstawie danych z próby. Statystyki próby (oznaczone jako b_0 i b_1) są obliczane jako oszacowania parametrów populacji β_0 i β_1. Zastąpienie wartości statystyk z próby b_0 i b_1 zamiast β_0 i β_1 w równaniu regresji daje nam nowe, oszacowane równanie regresji. Oszacowane równanie regresji dla prostej regresji liniowej jest następujące:

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

Wykres oszacowanej prostej regresji liniowej nazywany jest oszacowaną linią regresji. b_0 reprezentuje wartość początkową linii regresji, b_1 to współczynnik kierunkowy linii.

Poniżej pokazane jest, jak użyć metody najmniejszych kwadratów do obliczenia wartości b_0 i b_1 w oszacowanym równaniu regresji.

Ogólnie $\hat{y}$ (wynik dla $E(y)$) to średnia wartość y dla danej wartości x. Gdybyśmy teraz chcieli oszacować oczekiwaną wartość kwartalnej sprzedaży dla wszystkich restauracji Best Burger zlokalizowanych w pobliżu kampusów z 10000 studentów, wartość x zostałaby zastąpiona wartością 10000 w ostatnim równaniu. W niektórych przypadkach możemy być jednak bardziej zainteresowani prognozowaniem sprzedaży tylko dla jednej konkretnej restauracji. Załóżmy na przykład, że chcesz prognozować kwartalną sprzedaż dla restauracji, którą planujesz zbudować w pobliżu uczelni z 10000 studentów. Jak się okazuje, nawet w tym

przypadku najlepszym predyktorem (źródłem prognozy) wartości y dla danej wartości x jest wartość $\hat{y}$.

5.4 Metoda najmniejszych kwadratów

Metoda najmniejszych kwadratów to procedura, w której wykorzystując przykładowe dane, znajdujemy równanie szacowanej linii regresji. Aby zilustrować metodę najmniejszych kwadratów, załóżmy, że dane zostały zebrane z próby 10 restauracji Best Burger w pobliżu kampusów uniwersyteckich. Przez x_i oznaczymy wielkość populacji studentów (w tysiącach), a przez y_i wielkość kwartalnej sprzedaży (w tysiącach EUR). x_i w y_i dla 10 przykładowych restauracji podsumowano

w poniższej tabeli (rysunek 5.2). Widzimy, że restauracja 1, z x_1 = 2 i y_1 = 58, znajduje się w pobliżu kampusu z 2000 studentów i osiąga kwartalną sprzedaż w wysokości 58 000 EUR. Restauracja 2, z x_2= 6 i y_2 = 105, znajduje się w pobliżu kampusu z 6000 studentów i osiąga kwartalną sprzedaż w wysokości 105 000 €. Restauracja o najwyższej wartości sprzedaży to restauracja 10, która znajduje się w pobliżu kampusu z 26 000 studentów i osiąga kwartalną sprzedaż w wysokości 202 000 €.

Poniżej na rysunku 5.2. przedstawiono wykres punktowy danych nazywany też wykresem rozrzutu. Populacja studentów jest pokazana na osi poziomej, a kwartalna sprzedaż na osi pionowej. Wykresy rozrzutu dla analizy regresji są konstruowane ze zmienną niezależną x na osi poziomej i zmienną zależną y na osi pionowej. Wykres rozrzutu pozwala nam zatem wyciągnąć wstępne wnioski na temat możliwego związku między zmiennymi.

Restauracja	Populacja studentów (tys.)	Sprzedaż kwartalna (tys. €)
i	x_i	y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Rysunek 5.2 Wykres punktowy danych

Jakie wstępne wnioski można wyciągnąć z poniższego rysunku 5.3? Wyższa sprzedaż kwartalna występuje w kampusach z większą populacją studentów. Ponadto istnieje stała zależność między wielkością populacji studentów a kwartalną sprzedażą, którą można opisać linią prostą. Pomiędzy x i y istnieje dodatnia zależność liniowa. Dlatego wybrany został model prostej regresji liniowej, aby przedstawić związek między kwartalną sprzedażą a populacją studentów. Biorąc pod uwagę ten wybór, naszym następnym zadaniem jest wykorzystanie przykładowej tabeli danych do określenia wartości b_0 i b_1, które są ważnymi parametrami w estymacji równania prostej regresji liniowej. Dla i-tej restauracji oszacowane równanie regresji to:

$$\hat{y}_i = b_0 + b_1 x_i$$

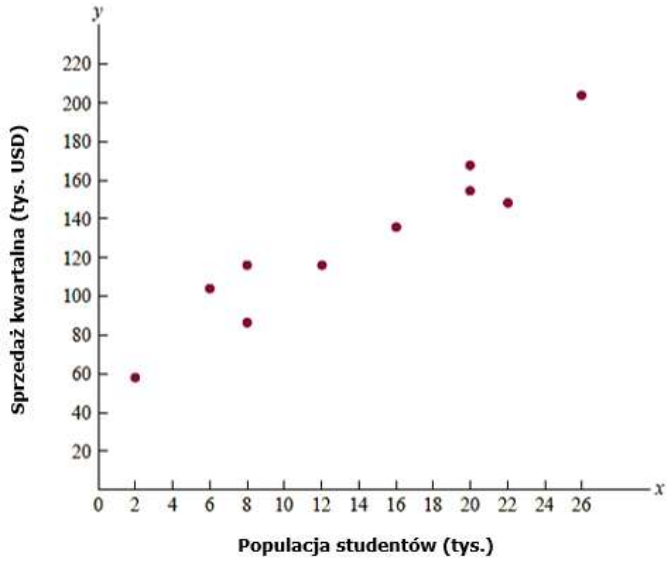
gdzie:

$\hat{y}_i$ - modelowa (teoretyczna) wartość sprzedaży kwartalnej (1000 EUR) dla i-tej restauracji,

b_0 - wyraz wolny oszacowanej linii regresji, inaczej przecięcie (z osią y),

b_1 - współczynnik kierunkowy oszacowanej linii regresji, inaczej nachylenie

x_i _ wielkość populacji studentów (1000) dla i-tej restauracji.



Rysunek 5.3 Wykres punktowy

y_i oznacza zaobserwowaną (rzeczywistą) sprzedaż dla restauracji i, a $\hat{y}_i$, teoretyczna wartość sprzedaży dla restauracji i, każda restauracja w próbie będzie miała zaobserwowaną wartość sprzedaży y_i i teoretyczną wartość sprzedaży $\hat{y}_i$. Aby oszacowana linia regresji zapewniała dobre dopasowanie do danych, chcemy, aby różnice między obserwowanymi wartościami sprzedaży a teoretycznymi wartościami sprzedaży były jak najmniejsze.

Metoda najmniejszych kwadratów wykorzystuje przykładowe dane do uzyskania wartości b_0 i b_1.

Minimalizacja sumy kwadratów odchyleń między obserwowanymi wartościami zmiennej zależnej y_i a przewidywaną wartością zmiennej zależnej $\hat{y}$. Punktem wyjścia do obliczenia minimalnej sumy metodą najmniejszych kwadratów jest wyrażenie:

Kryterium sumy minimalnej: $\min \sum(y_i - \hat{y}_i)^2$,

gdzie

y_i - zaobserwowana wartość zmiennej zależnej dla i-tej obserwacji,

$\hat{y}_i$ – teoretyczna wartość zmiennej zależnej dla i-tej obserwacji.

Współczynnik kierunkowy linii regresji i wyraz wolny:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

x_i _ wartość zmiennej niezależnej dla i-tej obserwacji,

y_i _ wartość zmiennej zależnej dla i-tej obserwacji,

$\bar{x}$ _ średnia wartość dla zmiennej niezależnej,

$\bar{y}$ _ średnia wartość dla zmiennej zależnej,

n _ całkowita liczba obserwacji.

Poniżej przedstawiono niektóre obliczenia potrzebne do opracowania szacunkowej linii regresji metodą najmniejszych kwadratów. Przy próbie 10 restauracji mamy n=10 obserwacji. Powyższe równania wymagają najpierw obliczenia średniej wartości x i średniej wartości y.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{140}{10} = 14, \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{1300}{10} = 130$$

Alternatywne równanie obliczeniowe b_1:

$$b_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

Korzystając z ostatnich równań i informacji na rysunku 5.4, można obliczyć współczynnik kierunkowy linii regresji dla przykładu restauracji Best Burger. Obliczenie nachylenia (b_1) przedstawiono poniżej.

Rysunek 5.4 przedstawia wykres tego równania na wykresie rozrzutu.

Nachylenie oszacowanego równania regresji lub współczynnik kierunkowy równania ($b_1 = 5$) jest dodatnie.

Restauracja i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Suma:	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

Rysunek 5.4 Wykres równania na wykresie rozrzutu.

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{2840}{568} = 5$$

Następnie obliczany jest wyraz wolny (b_0).

$$b_0 = \bar{y} - b_1\bar{x} = 130 - 5(14) = 60$$

W ten sposób szacowane jest równanie regresji:

$$\hat{y} = 60 + 5x$$

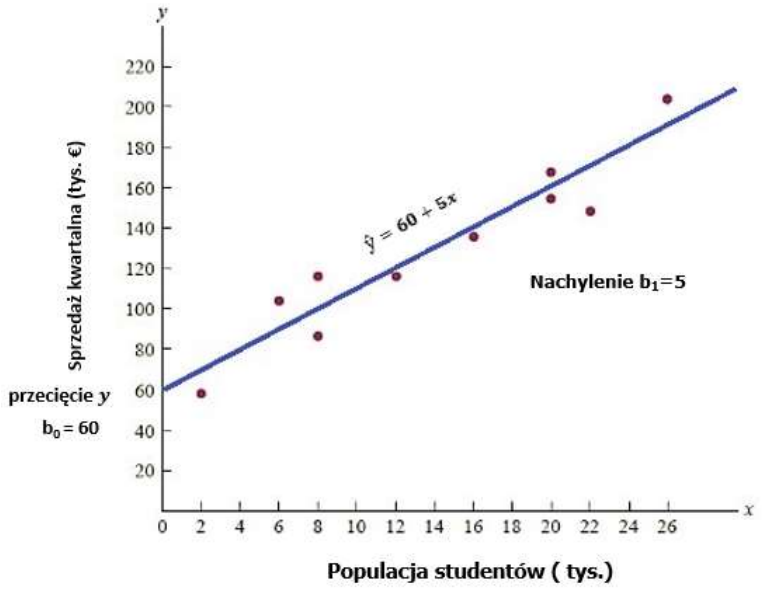
Rysunek 5.5 przedstawia wykres tego równania na wykresie punktowym.

Nachylenie oszacowanego równania regresji (b_1 = 5) jest dodatnie, co oznacza, że wraz ze wzrostem populacji studentów wzrasta sprzedaż. W rzeczywistości możemy wywnioskować (w oparciu o zmierzoną sprzedaż w tysiącach i populację studentów w tysiącach), co oznacza, że wzrost populacji studentów o 1000 wiąże się ze wzrostem oczekiwanej sprzedaży o 5000; tj. oczekuje się, że kwartalna sprzedaż wzrośnie o 5 USD na studenta.

Jeśli uważamy, że równanie regresji, oszacowane metodą najmniejszych kwadratów, odpowiednio opisuje związek między x i y, rozsądne wydaje się użycie oszacowanego równania regresji do przewidzenia wartości y dla danej wartości x. Na przykład, jeśli chcemy przewidzieć kwartalną sprzedaż dla restauracji znajdującej się w pobliżu kampusu z 16 000 studentów, obliczona byłaby ona na podstawie wzoru:

$$\hat{y} = 60 + 5 * 16 = 140$$

W związku z tym należy założyć, że kwartalna sprzedaż tej restauracji wyniesie 140 000. W kolejnych sekcjach omawiane są metody oceny stosowności wykorzystania oszacowanego równania regresji do estymacji i prognozowania.



Rysunek 5.5 Wykres regresji prostej kwartalnej sprzedaży

5.5 Współczynnik determinacji

Dla przykładu restauracji Best Burger opracowane zostało szacunkowe równanie regresji $y = 60 + 5x$ dla w przybliżeniu liniowej zależności między wielkością populacji studentów x a kwartalną sprzedażą y. Teraz pytanie brzmi: jak dobrze oszacowane równanie regresji pasuje do danych? W tej sekcji jest pokazane, że współczynnik determinacji zapewnia miarę dobroci (stopnia) dopasowania dla oszacowanego równania regresji. Dla i-tej obserwacji różnica między zaobserwowaną wartością zmiennej zależnej y_i a przewidywaną wartością zmiennej zależnej nazywana jest i - tą wartością rezydualną. Suma kwadratów tych reszt lub błędów jest funkcją nachylenia oraz wyrazu wolnego. Metodą najmniejszych kwadratów wyznaczamy wartości b_1 i b_0, dla których powyższa funkcja osiąga minimum. Wielkość ta oznaczana jest jako SSE.

$$SSE = \sum (y_i - \hat{y}_i)^2$$

Wartość SSE jest miarą błędu w wykorzystaniu oszacowanego równania regresji do opisu wartości zmiennej zależnej w próbie. Rysunek 5.6 pokazuje obliczenia potrzebne do obliczenia sumy kwadratów błędów dla przypadku Best Burger.

Restauracja i	x_i=Populacja studentów (tys.)	y_i=Sprzedaż kwartalna (tys. €)	Prognozowana sprzedaż $\hat{y}_i = 60 + 5x$	Błąd $y_i - \hat{y}_i$	Kwadrat błędu $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
					SSE = 1530

Rysunek 5.6 Kwadraty błędów dla przypadku Best Burger

Załóżmy, że poproszono nas o oszacowanie kwartalnej sprzedaży bez znajomości wielkości populacji studentów. Nie znając żadnych powiązanych zmiennych, użylibyśmy średniej z próby jako oszacowania kwartalnej sprzedaży w dowolnej restauracji. Tabela na rysunku 5.6 pokazuje, że dla danych sprzedaży $\sum y_i$ = 1300. Dlatego średnia kwartalna wartość sprzedaży dla próby 10 restauracji Best Burger wynosi $\sum y_i$/n = 1300/10 = 130. Na rysunku 5.7 pokazana jest suma kwadratów odchyleń uzyskanych przy użyciu średniej próby 130 do przewidywania wartości kwartalnej sprzedaży dla każdej restauracji w próbie. Dla i-tej restauracji w próbie różnica $y_i - \hat{y}_i$ stanowi miarę błędu, który jest uwzględniony w aplikacji modelu do wielkości sprzedaży. Odpowiednia suma kwadratów, zwana całkowitą sumą kwadratów, jest oznaczana przez SST.

$$SST = \sum (y_i - \bar{y})^2$$

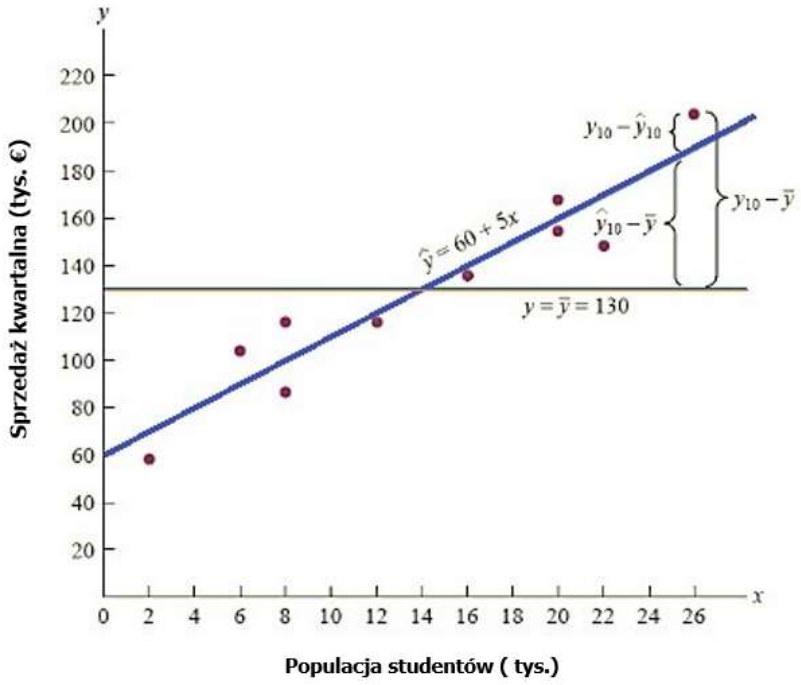
Restauracja i	x_i=Populacja studentów (tys.)	y_i=Sprzedaż kwartalna (tys. €)	Odchylenie $y_i - \bar{y}$	Kwadrat odchylenia $(y_i - \bar{y})^2$
1	2	58	−72	5184
2	6	105	−25	625
3	8	88	−42	1764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1521
9	22	149	19	361
10	26	202	72	5184
				SST = 15,730

Rysunek 5.7 Suma kwadratów

Suma na dole ostatniej kolumny na rysunku 5.7 to całkowita suma kwadratów dla restauracji BestBurger SST = 15 730. Na rysunku 5.8 pokazana jest szacowana linia regresji $y = 60 + 5x$ i linia odpowiadająca y = 130. Należy zauważyć, że punkty skupiają się bardziej wokół szacowanej linii regresji niż wokół linii y = 130. Na przykład dla 10. restauracji w próbie widzimy, że błąd jest znacznie większy, gdy 130 jest używana jako wielkość y, niż gdy użyjemy równania $y = 60 + 5x$, dla którego wynosi on 190. Można myśleć o SST jako o pomiarze tego, jak dobrze obserwacje skupiają się wokół linii o stałej wartości y wynoszącej 130, a SSE jako o pomiarze tego, jak dobrze obserwacje skupiają się wokół prostej regresji.

Aby zmierzyć, jak bardzo wartości na oszacowanej linii regresji odbiegają od średniej, obliczana jest kolejna suma kwadratów. Ta suma kwadratów, zwana sumą kwadratów z powodu regresji, jest oznaczana jako SSR.

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$



Rysunek 5.8 Linia regresji dla przypadku Best Burger

Z poprzedniej dyskusji powinniśmy się spodziewać, że SST, SSR i SSE są ze sobą powiązane. W rzeczywistości związek między tymi trzema sumami kwadratów jest jednym z najważniejszych wyników w statystyce.

5.6 Związek między SST, SSR i SSE

$$SST = SSR + SSE$$

gdzie:

SST = całkowita suma kwadratów,

SSR = suma kwadratów wynikająca z regresji,

SSE = suma kwadratów spowodowana błędem.

Równanie $(SST = SSR + SSE)$ pokazuje, że całkowitą sumę kwadratów można podzielić na dwa składniki, sumę kwadratów wielkości teoretycznych regresji i sumę kwadratów błędów. Jeśli więc znane są wartości dowolnych dwóch z tych sum kwadratów, trzecią sumę kwadratów można łatwo obliczyć. Na przykład w przypadku restauracji Best Burger wiadomo

już, że SSE = 1530 i SST = 15730; dlatego rozwiązując SSR w powyższym równaniu, okazuje się, że suma kwadratów wynikająca z regresji wynosi:

$$SSR = SST - SSE = 15730 - 1530 = 14200$$

Zobaczmy teraz, jak możemy wykorzystać trzy sumy kwadratów, SST, SSR i SSE, aby zapewnić kryterium dobrego dopasowania dla oszacowanego równania regresji. Oszacowane równanie regresji zapewniłoby idealne dopasowanie, gdyby każda wartość zmiennej zależnej $yi\_$ leżała losowo na oszacowanej linii regresji. W tym przypadku wynosiłaby ona zero dla każdej obserwacji, co skutkowałoby SSE =0. Ponieważ SST = SSR + SSE, widzimy, że dla idealnego dopasowania SSR musi być równe SST, a stosunek (SSR/SST) musi być równy jeden. Gorsze dopasowanie spowoduje większe wartości SSE. Rozwiązując równanie dla SSE, widzimy, że SSE = SST - SSR. Dlatego największa wartość SSE (a tym samym najgorsze dopasowanie) występuje, gdy SSR = 0 i SSE = SST.

Stosunek SSR/SST jest wykorzystywany do estymacji, która ma wartości od zera do jednego pasujące do szacowanego równania regresji.

Współczynnik ten nazywany jest współczynnikiem determinacji i jest oznaczany przez r^2.

$$r^2 = \frac{SSR}{SST}$$

Dla przykładu restauracji Best Burger wartość współczynnika determinacji wynosi:

$$r^2 = \frac{SSR}{SST} = \frac{14200}{15730} = 0{,}9027$$

Gdy współczynnik determinacji jest wyrażony w procentach, r^2 można go interpretować jako procent całkowitej sumy kwadratów wyjaśniony za pomocą oszacowanego równania regresji. W przypadku Best Burger Restaurants możemy stwierdzić, że 90,27% całkowitej sumy kwadratów można wyjaśnić za pomocą oszacowanego równania regresji $y = 60 + 5x$, aby przewidzieć kwartalną sprzedaż. Innymi słowy, 90,27% zmienności sprzedaży można wyjaśnić liniową zależnością między wielkością populacji studentów a sprzedażą. Powinniśmy być zadowoleni, że tak dobrze pasuje do oszacowanego równania regresji.

5.7 Współczynnik korelacji

Współczynnik korelacji można traktować jako opisową miarę siły współzależności liniowej między dwiema zmiennymi, x i y. Wartości współczynnika korelacji zawsze mieszczą się

w przedziale od -1 do +1. Wartość +1 oznacza, że dwie zmienne x i y są doskonale skorelowane. Oznacza to, że linia prosta o dodatnim nachyleniu stanowi wykres punktowy obserwacji. Dlatego wykres rozrzutu nazywa się też wykresem korelacji. Wartość -1 oznacza, że zmienne x i y są doskonale skorelowane. Wszystkie punkty danych leżą na prostej o ujemnym nachyleniu. Wartości współczynnika korelacji bliskie zeru oznaczają, że x i y nie współzależą liniowo.

Jeśli przeprowadzono już analizę regresji i obliczono współczynnik determinacji r^2, współczynnik korelacji próby można obliczyć w następujący sposób.

$$r_{xy} = (znak\ b_1)\sqrt{współczynnik\ determinacji} \quad (13)$$

$$r_{xy} = (znak\ b_1)\sqrt{r^2} \quad (14)$$

WSPÓŁCZYNNIK KORELACJI PEARSONA: PRZYKŁADOWE DANE

$$r_{xy} = \frac{V_{xy}}{V_x V_y} = \frac{\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}}{\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$$

$$r_{xy} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

gdzie:

$V_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}$, $V_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$, $V_y = \sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}$

Znak współczynnika korelacji próby jest dodatni, jeśli oszacowane równanie regresji ma dodatnie nachylenie ($b_1 > 0$) i ujemny, jeśli oszacowane równanie regresji ma ujemne nachylenie ($b_1 < 0$).

W przypadku Best Burger wartość współczynnika determinacji odpowiadającego oszacowanemu równaniu regresji $y = 60 + 5x$ wynosi 0,9027. Ponieważ nachylenie oszacowanego równania regresji jest dodatnie, równanie (13-14) pokazuje, że próbny współczynnik korelacji wynosi r_{xy}=0,9501. Na podstawie próbnego współczynnika korelacji można wywnioskować, że istnieje silna dodatnia zależność liniowa między x a y.

W przypadku liniowej zależności między dwiema zmiennymi, zarówno współczynniki determinacji, jak i współczynnik korelacji stanowią miarę siły związku w próbce.

Współczynnik determinacji zapewnia miarę między zero a jeden, podczas gdy współczynnik korelacji zapewnia miarę między -1 a +1. Chociaż współczynnik korelacji jest ograniczony do liniowej zależności między dwiema zmiennymi, współczynnik determinacji może być stosowany do zależności nieliniowych i do zależności, które mają dwie lub więcej niezależnych zmiennych. W ten sposób współczynnik determinacji zapewnia szerszy zakres zastosowań.

5.8 Model regresji wielorakiej

W tej sekcji kontynuujemy nasze badanie analizy regresji, rozważając sytuacje obejmujące dwie lub więcej zmiennych niezależnych. Ten obszar tematyczny, zwany analizą regresji wielorakiej, pozwala nam wziąć pod uwagę więcej czynników, a tym samym uzyskać lepsze prognozy niż jest to możliwe w przypadku jednorakiej regresji liniowej.

Analiza regresji wielorakiej to badanie, w jaki sposób zmienna zależna y jest powiązana z dwiema lub więcej zmiennymi niezależnymi. W ogólnym przypadku oznaczymy przez p liczbę zmiennych niezależnych.

5.9 Model i równanie regresji

Pojęcia modelu regresji i równania regresji wprowadzone w poprzedniej sekcji mają zastosowanie w przypadku regresji wielorakiej. Równanie opisujące, w jaki sposób zmienna zależna y jest powiązana ze zmiennymi niezależnymi $x_1, x_2, \ldots, x_p$ i wielkością błędu, nazywane jest modelem regresji wielorakiej. Zaczynamy od założenia, że model regresji wielorakiej ma następującą postać.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

W modelu regresji wielorakiej $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ są parametrami, a składnik błędu (ϵ) jest zmienną losową. Dokładna analiza tego modelu ujawnia, że y jest liniową funkcją zmiennych $x_1, x_2, \ldots, x_p$ plus składnik błędu ϵ epsilon. Wielkość błędu uwzględnia zmienność y, której nie można wyjaśnić liniowym efektem p niezależnych zmiennych.

W sekcji 5.10 omawiane są założenia dla modelu regresji wielorakiej i epsilon. Jednym z założeń jest to, że średnia lub wartość oczekiwana (ϵ) wynosi zero. Konsekwencją tego założenia jest to, że średnia lub wartość oczekiwana y, oznaczana jako $E(y)$, jest równa $\beta_0 +$

$\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$. Równanie opisujące zależność wartości średniej y od $x_1, x_2, \ldots, x_p$ nazywane jest równaniem regresji wielorakiej.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p (5.9)$$

5.10 Oszacowane równanie regresji wielorakiej

Jeśli znane są wartości $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, równanie (5.9), można wykorzystać do obliczenia średniej wartości y przy danych wartościach $x_1, x_2, \ldots, x_p$. Niestety, te wartości parametrów zazwyczaj nie są znane i muszą zostać oszacowane na podstawie danych z próby. Prosta próba losowa jest używana do obliczenia statystyk próby $b_0, b_1, b_2, \ldots, b_p$, które będą używane jako estymatory punktowe parametrów $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. Te przykładowe statystyki zapewniają następujące oszacowanie równania regresji wielokrotnej:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

gdzie:

$b_0, b_1, b_2, \ldots, b_p$ są oszacowaniami $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$,

$\hat{y} =$ przewidywana wartość zmiennej zależnej.

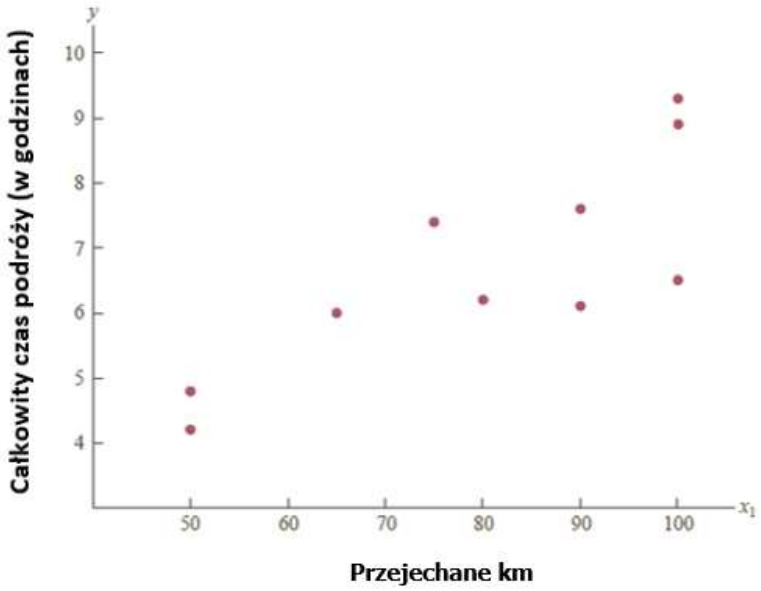
Przykład: Frigo Transport Company

Jako ilustrację analizy regresji wielorakiej rozważymy problem, przed którym stoi Frigo Trucking Company, niezależna firma transportowa w południowych Włoszech. Największa część działalności Frigo obejmuje dostawy na całym obszarze lokalnym. Aby lepiej opracować harmonogramy pracy, menedżerowie chcą zaplanować wspólny dzienny czas podróży dla swoich kierowców.

Początkowo menedżerowie uważali, że całkowity dzienny czas podróży będzie ściśle powiązany z liczbą kilometrów przejechanych w codziennych dostawach. Prosta losowa próba 10 przydziałów kierowców dostarczyła danych pokazanych w tabeli 5.1 i wykresu rozrzutu (rys.5.9). Po zapoznaniu się z tym wykresem rozrzutu menedżerowie założyli, że model jednorakiej regresji liniowej może być użyty do opisania zależności $y = \beta_0 + \beta_1 x_1 + \epsilon$ między całkowitym czasem podróży (y) a liczbą przejechanych kilometrów (x_1).

Tabela 5.1 Dane dla przykładowej firmy Frigo Transport Company

Zadanie kierowcy	x_i =Przejechane km	y=Czas podróży (w godzinach)
1	100	9.3
2	50	4.8
3	100	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1



Rysunek 5.9 Wykres rozrzutu dla Frigo Transport Company.

Do oszacowania parametrów β_0 i β_1 wykorzystano metodę najmniejszych kwadratów w celu opracowania szacunkowej regresji:

$$\hat{y} = b_0 + b_1 x_1$$

Powyższy rysunek przedstawia wyniki programu Minitab przy użyciu prostej regresji liniowej dla danych z powyższej tabeli. Oszacowane równanie regresji to:

$$\hat{y} = 1{,}27 + 0{,}0678x_1$$

Wartość *F* równa 15,81 i odpowiadająca jej wartość *p-value* równa 0,004 wskazują, że związek jest istotny na poziomie istotności 0,05. Oznacza to, że można odrzucić $H_0: \beta_1 = 0$, ponieważ wartość *p-value* jest mniejsza niż poziom istotności alfa (α=0,05). Należy zauważyć, że ten sam wniosek wynika z wartości t = 3,98 i związanej z nią wartości *p-value* równej 0,004. Możemy zatem stwierdzić, że związek między całkowitym czasem podróży a liczbą przejechanych kilometrów jest znaczący. Dłuższy czas podróży wiąże się z większą liczbą przejechanych kilometrów i odwrotnie. Przy współczynniku determinacji (wyrażonym w procentach) $r^2 = 66{,}4\,\%$ widać, że 66,4% zmienności czasu podróży można wyjaśnić liniowym efektem liczby przejechanych kilometrów.

Wynik ten jest całkiem dobry, ale menedżerowie mogą rozważyć dodanie drugiej zmiennej niezależnej, aby zmniejszyć błąd regresji (błąd szczątkowy) dla zmiennej zależnej.

DANE WYJŚCIOWE MINITAB DLA BUTLER TRUCKING
Z JEDNĄ ZMIENNĄ NIEZALEŻNĄ

Równanie regresji ma postać:
Czas=1,27+0,0678 km

Predyktor	Współczynnik	Współczynnik SE	T	p-value
Stała	1,274	1,401	0,91	0,39
Nachylenie (b_1)	0,06783	0,01706	3,98	0,004

S=1,00179 R-Sq=66,4% R-SQ(adj)=62,2%

Analiza wariancji

Źrodło	DF	SS	MS	F	p-value
Regresja	1	15,871	15,871	15,81	0,004
Błąd szczątkowy	8	8,029	1,004		
Suma	9	23,9			

Rysunek 5.10 Wyniki z jedną zmienną niezależną

Próbując zidentyfikować drugą zmienną niezależną, menedżerowie uznali, że liczba dostaw może również przyczynić się do całkowitego czasu podróży. Dane Frigo Trucking: liczbę

przejechanych kilometrów (x_1) i liczbę dostaw (x_2), jako zmienne niezależne, przedstawiono na rysunku 5.11. Oszacowane równanie regresji to:

$$\hat{y} = -0{,}869 + 0{,}0611x_1 + 0{,}923x_2$$

DANE DLA FRIGO TRUCKING Z PRZEJECHANYMI KILOMETRAMI (x_1) I DOSTAWAMI (x_2) JAKO ZMIENNYMI NIEZALEŻNYMI

Zadanie kierowcy	x_1 =Przejechane km	x_2 =Liczba dostaw	y=Czas podróży (w godzinach)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

Rysunek 5.11 Dane Frigo Trucing i zmienne niezależne

Przyjrzyjmy się bliżej wartościom b_1= 0,0611 i b_2= 0,923 w ostatnim równaniu.

Uwaga dotycząca interpretacji współczynników

W tym momencie można poczynić jedną uwagę na temat związku między oszacowanym równaniem regresji prostej, z tylko przejechanymi kilometrami jako zmienną niezależną, a równaniem uwzględniającym również liczbę dostaw jako drugą zmienną niezależną. Wartość b_1 nie jest taka sama w obu przypadkach. W prostej regresji liniowej interpretujemy b_1 jako oszacowanie zmiany y dla zmiany zmiennej niezależnej o jedną jednostkę. W analizie regresji wielorakiej interpretacja ta musi zostać nieco zmodyfikowana. Oznacza to, że w analizie regresji wielorakiej każdy współczynnik regresji jest interpretowany w sposób, który reprezentowałby szacunkową zmianę y odpowiadającą zmianie x_i o jedną jednostkę, gdy wszystkie inne zmienne niezależne są utrzymywane na stałym poziomie.

W przypadku Frigo Trucking obejmuje to dwie niezależne zmienne, b_1=0,0611 i b_2= 0,923.

DANE WYJŚCIOWE MINITAB DLA FRIGO TRUCKING Z DWIEMA ZMIENNYMI NIEZALEŻNYMI

Równanie regresji ma postać:
Czas=-0,869+0,0611 km + 0,923 Dostaw

Predyktor	Współczynnik	Współczynnik SE	T	p-value
Stała	-0,8687	0,9515	-0,91	0,392
Nachylenie (b_1)	0,061135	0,009888	6,18	0,000
Dostawy	0,9234	0,2211	4,18	0,004

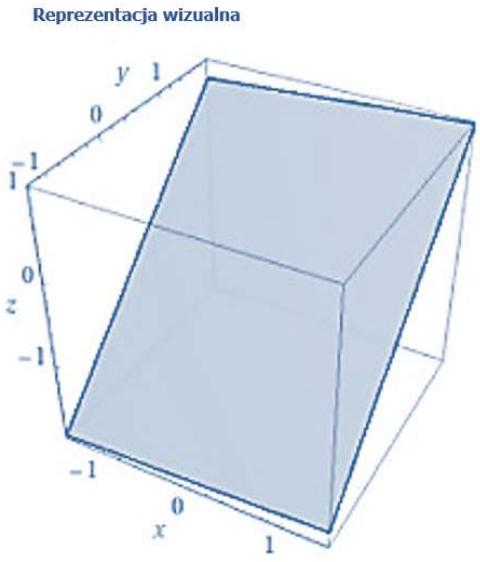
S=0,573142 R-Sq=90,4% R-SQ(adj)=87,6%

Analiza wariancji

Źrodło	DF	SS	MS	F	p-value
Regresja	2	21,601	10,8	32,88	0,000
Błąd szczątkowy	7	2,299	0,328		
Suma	9	23,900			

Rysunek 5.12 Wyniki dla Frigo Trucking z dwiema zmiennymi niezależnymi

Zatem 0,0611 godziny to szacunkowy oczekiwany wzrost czasu podróży odpowiadający wzrostowi o jeden km na przejechaną odległość, gdy liczba dostaw jest stała. Podobnie, ponieważ b_2 =0,923, szacowany oczekiwany wzrost czasu podróży odpowiadający wzrostowi o jedną dostawę, gdy liczba przejechanych kilometrów jest stała, wynosi 0,923 godziny.



Rysunek 5.13 Wizualna reprezentacja wyników dla przypadku Frigo Trucking

BIBLIOGRAFIA DO ROZDZIAŁU 5.

- Introductory Statistics. Bentham Science Publishers, Kahl, A. (Publish 2023). DOI:10.2174/9789815123135123010l
- Introductory Statistics 2e, Openstax, Rice University, Houston, Texas 77005, Jun 23, senior contributing authors: Barbara Illowsky and Susan dean, De anza college, Publish Date: Dec 13, 2023, (https://openstax.org/details/books/introductory-statistics-2e);
- Introductory Statistics 4th Edition, Susan Dean and Barbara Illowsky, Adapted by Riyanti Boyd & Natalia Casper (Published 2013 by OpenStax College) July 2021, (http://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf);
- Journal of the Royal Statistical Society 2024**,** A reputable journal publishing cutting-edge research and articles on various aspects of statistics, including theoretical advancements and practical applications. Recent issues have featured studies on sampling and hypothesis testing.
- Introductory Statistics 7th Edition, Prem S. Mann, eastern Connecticut state university with the help of Christopher Jay Lacke, Rowan university, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, 2011
- Introduction to statistics, made easy second edition, Prof. Dr. Hamid Al-Oklah Dr. Said Titi Mr. Tareq Alodat, March 2014
- Statistics for Business and Economics, Thirteenth Edition, David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, 2017, 2015 Cengage Learning®
- Statistics for Business, First edition, Derek L Waller, 2008 Copyright © 2008, Derek L Waller, Published by Elsevier Inc. All rights reserved