Instytut Sterowania i Systemów Informatycznych
Uniwersytet Zielonogórski

Podstawy analityki biznesowej

Ćwiczenie 4: Korelacja i regresja liniowa

Zadanie 1. Plik **resting.sas7bdat** zawiera wzrost (w cm) i tętno spoczynkowe (w uderzeniach na minutę) dla pewnej grupy pacjentów szpitala. Zależy nam na określeniu czy występuje zależność między tymi dwiema zmiennymi.

(a) Zaimportować plik **resting.sas7bdat** do biblioteki **sasue** utworzonej na poprzednich zajęciach i wyświetlić jego zawartość.

(b) Najpierw utworzymy tzw. wykres punktowy. W tym celu należy otworzyć **Zadania ►Wykresy ►Wykres punktowy**. W **Dane ►Dane** wybrać **sasue.resting**. W **Dane ►Role ►Zmienna x** dodać **height**, w **Dane ►Role ►Zmienna y** natomiast – **pulse**. W **Opcje ►Oś X** odznaczyć **Pokaż linie siatki**. To samo zrobić dla drugiej osi, po czym kliknąć **Uruchom**. Jaką tendencję ma tętno, gdy wzrost pacjenta staje się coraz większy?

(c) Miarą zależności liniowej między dwiema zmiennymi jest *współczynnik korelacji liniowej Pearsona r*. Jego wartość uzyskujemy otwierając **Zadania ►Statystyka ►Analiza korelacji**. W **Dane ►Dane** wybrać **sasue.resting**. W **Dane ►Role ►Zmienne analizowane** dodać **height** oraz **pulse**. W **Opcje ►Statystyki ►Wyświetl statystyki** wybrać **Wybrane statystyki**. Kliknąć **Uruchom**. Wartością współczynnika korelacji między wysokością i tętnem jest $r=0.22$, co wskazuje na stosunkowo słaby dodatni związek między zmiennymi. Ta wartość jest oceną odpowiedniej wartości w populacji, co do której można przeprowadzać różnego rodzaju testy. Najczęściej weryfikowaną hipotezą jest zerowość wartości współczynnika korelacji w populacji (jest to równoznaczne z brakiem liniowej zależności między zmiennymi). Przy jej prawdziwości statystyka testowa ma postać

$$t=r\sqrt{\frac{n-2}{1-r^2}},$$

gdzie: n – wielkość próby. Powinna mieć ona rozkład t Studenta o $n-2$ stopniach swobody. SAS Studio wyświetla rezultat testu pod nazwą **Prawd. > |t| przy H0: rho = 0**, z odpowiednią p-wartością pod zapisaną pod wartością r. Ile wynosi ona w rozważanym przypadku i co to oznacza?

Powyższy test domyślnie zakłada, że dwie zmienne losowe są opisywane łącznym rozkładem normalnym, jednak w praktyce spełnienie tego założenia bada się rzadko.

(d) Od określenia stopnia skorelowania między zmiennymi, często jesteśmy bardziej zainteresowani równaniem wiążącym jedną zmienną z drugą, którego można byłoby użyć do przewidywania wartości jednej ze zmiennych na podstawie wartości drugiej. Najczęstszym modelem jest *równanie prostej regresji liniowej*:

$$y_i=\alpha+\beta x_i+\varepsilon_i,$$

w którym x_i, y_i reprezentują wartości dwóch zmiennych (nazywanych odpowiednio zmienną objaśniającą i odpowiedzią) dla i-tej obserwacji, a ε_i jest wartością błędu (tzn. różnicą między zaobserwowaną wartością y_i i tą samą wartością przewidywaną przez model, czyli $\alpha+\beta x_i$). Formuły do obliczania ocen współczynników α, β zostaną

podane na wykładzie. O błędach zakłada się, że mają rozkład normalny o stałej wariancji.

(e) Dopasowanie modelu prostej regresji liniowej do danych o wzroście i tętnie wraz z wykresem uzyskuje się otwierając **Zadania ►Statystyka ►Regresja liniowa**. W **Dane ►Dane** wybrać **sasue.resting**. W **Dane ►Role ►Zmienna zależna** dodać **pulse**. W **Dane ►Role ►Zmienne ciągłe** dodać **height**. W **Model ►Efekty modelu** wybrać **height** i kliknąć przycisk **Dodaj**. W **Opcje ►Wykresy ►Wykresy punktowe** wybrać **Wykres dopasowania dla pojedynczej zmiennej ciągłej** i odznaczyć pozostałe wykresy. Kliknąć **Uruchom**.

Tabela **Analiza wariancji** określa jak zmienność zmiennej *y* rozdziela się na część spowodowaną dopasowywanym modelem oraz część spowodowaną występowaniem członu błędu (to zagadnienie będzie tematem następnych zajęć ćwiczeniowych). Związany z nią test *F* bada hipotezę o zerowości współczynnika nachylenia ($H_0: \beta = 0$). Jaka jest *p*-wartość tego testu i o co na tej podstawie można wywnioskować? Zwrócić uwagę na to, że *p*-wartość jest identyczna z wartością otrzymaną wcześniej dla współczynnika korelacji.

Najważniejszą wartością jest **R-kwadrat**, która jest kwadratem współczynnika korelacji *r* pomiędzy zaobserwowanymi wartościami odpowiedzi (czyli *y*) oraz wartościami odpowiedzi przewidywanymi przez dopasowany model. **R-kwadrat** określa wariancję odpowiedzi *y*, która jest objaśniana przez zmienną *x*. Z tabeli wynika, że tylko 5% wariancji tętna jest objaśniana przez wzrost.

Tabela **Oceny parametrów** zawiera obliczone oszacowania wyrazu wolnego α (wiersz **Intercept**) oraz współczynnika nachylenia β (wiersz **height**). Zgodnie z otrzymanym modelem, o jaką wartość wrośnie tętno jeśli wzrost zwiększy się o 1 cm? Wyznaczyć 95% przedział ufności dla oceny współczynnika nachylenia (*Wskazówka:* kolumna **Błąd standardowy** zawiera odchylenie standardowe oceny, więc wystarczy wykorzystać *regułę dwóch sigm*). Czy ten przedział zawiera wartość zero? Co to oznacza?

(f) Wykres dopasowania zawiera wykres punktowy z dopasowaną linią prostą oraz 95% granice ufności dla linii prostej odpowiadającej całej populacji pokazane jako obszar zacieniowany na jasnoniebiesko (szerokość przedziału ufności rośnie z odchyleniem od punktu środkowego prostej regresji). Na rysunku nakreślono również linie przerywane, będące 95% granicami przedziałów ufności dla przewidywanych wartości tętna. Przewidywane wartości oblicza się poprzez zastosowanie dopasowanego modelu, tzn.

$$\text{przewidywana wartość tętna} = \text{ocena wyrazu wolnego} + \text{ocena współczynnika nachylenia} \times \text{wysokość}$$

Zauważyć, że linia pozioma (tzn. z zerowym współczynnikiem nachylenia – co to oznacza?) może być łatwo wpasowana w obszar ufności dla populacyjnej linii prostej.

Zadanie 2. Zbiór danych **anaerob** zawiera dane zebrane w eksperymencie kinezjologicznym (kinezjologia zajmuje się psychologicznymi i fizjologicznymi reakcjami ludzkiego organizmu na krótkotrwały i bardzo intensywny wysiłek fizyczny). Badany człowiek wykonywał standardowe ćwiczenie fizyczne stopniowo zwiększając jego stopień trudności. Zbiór danych zawiera pary składające się z wartości absorpcji tlenu (zmienna **o2in**) oraz ilości wydychanego powietrza (zmienna **airout**).

(a) Narysować wykres punktowy ze zmienną **o2in** na osi odciętych oraz zmienną **airout** na osi rzędnych. Czy wskazuje on na korelację tych zmiennych? Jak silną? Jakiego typu (liniowa czy nieliniowa)?

(b) Obliczyć współczynnik korelacji liniowej. Co można powiedzieć o hipotezie o jego zerowości?

(c) Pomimo bardzo dużej wartości współczynnika korelacji liniowej, wykres punktowy wskazuje bardziej na nieliniowy charakter zależności między dwiema rozważanymi zmiennymi. Spróbujmy więc dopasować poniższy model, zawierajacy kwadratowy efekt absorpcji tlenu:

$$y_i = \alpha + \beta x_i + \beta_2 x_i^2 + \varepsilon_i.$$

Zauważmy, że mimo nieliniowości względem absorpcji tlenu, odpowiedź zależy liniowo od parametrów α, β_1, β_2. Ta własność istotnie ułatwia wnioskowanie. Przykładem w pełni nieliniowego modelu jest

$$y_i = \alpha_1 \exp(\beta x_i) + \alpha_2 \exp(\beta_2 x_i^2) + \varepsilon_i.$$

Analiza takich modeli jest jednak nieco złożona, a odpowiednie informacje można znaleźć w podręcznikach pod hasłem *regresji nieliniowej*.

Otworzyć **Zadania ►Statystyka ►Regresja liniowa**. W **Dane ►Dane** wybrać **sasue.anaerob**. W **Dane ►Role ►Zmienna zależna** dodać **airout**. W **Dane ►Role ►Zmienne ciągłe** dodać **o2in**. W **Model ►Efekty modelu** wybrać **o2in** i kliknąć przycisk **Dodaj**, a następnie kliknąć przycisk **Rząd wielomianu = n**. Pojawi się okienko z wartością **2** jako domyślną. Proszę kliknąć **Dodaj**. Spowoduje to dodanie członu **o2in * o2in** do efektów modelu. W **Opcje ►Wykresy ►Wykresy punktowe** wybrać **Wykres dopasowania dla pojedynczej zmiennej ciągłej** i odznaczyć pozostałe wykresy. Kliknąć **Uruchom**.

Czy wszystkie parametry dopasowanej funkcji kwadratowej można ocenić jako niezerowe?

Zadanie 3. Analizy regresji nie można zakończyć bez próby sprawdzenia przyjętych założeń o stałości wariancji i normalności błędu w modelu. Odchylenia od jednego lub obu z nich mogą podważać ewentualne konkluzje oparte o analizę regresji. Sprawdzenie założeń wiąże się przede wszystkim z użyciem tego lub innego typu reszt, z których najprostsza jest po prostu różnicą między zaobserowaną wartością *y* i wartością przewidywaną przez dopasowany model. Oto przykłady typowych wykresów diagnostycznych:

Reszty względem wartości przewidywanych: Jeśli dopasowany model jest adekwatny, narysowane punkty powinny wypełniać poziomy pas wokół wartości zerowej na osi odciętych. Odchylenia od tego stanu rzeczy mogą wskazywać, że postać funkcyjna zakładanego modelu jest niepoprawna lub, alternatywne, że wariancja nie jest stała.

Reszty względem wartości zmiennej x: ewentualny wzorzec, jaki tworzą punkty, może wskazywać na odchylenia od założenia o stałej wariancji lub niewłaściwą postać modelu.

Wykres kwantylowy reszt: używa się go do sprawdzenia założenia o rozkładzie normalnym błędów.

(a) Powtórzyć czynności z Zadania 1 dotyczącego zestawu danych **resting**, jednak przed kliknięciem **Uruchom** w **Opcje ►Wykresy ►Wykresy diagnostyczne i resztowe** wybrać **Wykresy diagnostyczne** i **Reszty dla każdej ze zmiennych objaśniających**. Poniżej, w **Wyświetl jako**, można wybrać **Panel wykresów** (wystarcza w większości sytuacji) lub **Pojedyncze wykresy** (przydatne, gdy interesują nas szczegóły wykresów). Kliknąć **Uruchom**.

(b) Przeanalizować wykresy **Reszta-Wartość przewidywana**, **Residual-Kwantyl** oraz **Procent-Reszta**. Co sobą reprezentują i o czym mówią?

(c) Odnaleźć wykres **Reszta-Height**. Czy daje on powód do zakwestionowania przyjętych założeń?

(d) Powtórzyć wszystkie czynności dla danych **anaerob**.

Zadanie 4. Zbiór danych **usbirths** zawiera liczbę noworodków w USA na tysiąc mieszkańców w latach 1940–1948. Ćwiczenie ilustruje, jak ważny w analizie wzorców pojawiających się na wykresach jest współczynnik kształtu, będacy stosunkiem fizycznej długości osi pionowej do długości osi poziomej.

(a) Otworzyć **Zadania ►Wykresy ►Wykres punktowy**. W **Dane ►Dane** wybrać **sasue.usbirth**. W **Dane ►Role ►Zmienna x** dodać **obsdate**. W **Dane ►Role ►Zmienna y** dodać **rate**. W **Opcje ►Oś x** odznaczyć **Pokaż linie siatki**. To samo zrobić dla osi *y*. W **Opcje ►Rozmiar wykresu** wymiary domyślne to 4.8×6.4 cali, czyli współczynnik kształtu 3:4. Kliknąć **Uruchom**. Co można powiedzieć o tendencjach w okresach 1940–1943, 1943–1946, rok 1946, 1947–1948? Jak powiązać to z zaangażowaniem USA w wojnie latach 1942–1945?

(b) Powtórzyć poprzedni punkt, zmieniając jednak w **Opcje ►Rozmiar wykresu** wymiary na szerokość 6 cali i wysokość 2 cale. Zauważyć istotne zwiększenie liczby szczegółów.

(c) Teraz spróbujemy uzyskać jeszcze więcej szczegółów ograniczając horyzont czasowy do lat 1940–1943. W **Dane ►Filtr warunków WHERE** wybrać **Zastosuj warunek WHERE**, a następnie w oknie **Łańcuch WHERE:** wpisać **year < 1944**. Ponownie uruchomić zadanie. Zauważyć cykle w ramach jednego roku. Najniższy wskaźnik urodzin występuje na początku lata, a najwyższy – jesienią.

(d) Powyższa obserwacja będzie jeszcze bardziej widoczna jeśli punkty połączymy linią. Otworzyć **Zadania ►Wykresy ►Wykres szeregu**. W **Dane ►Dane** wybrać **sasue.usbirth**. W **Dane ►Role ►Zmienna x** dodać **obsdate**. W **Dane ►Role ►Zmienna y** dodać **rate**. W **Dane ►Filtr warunków WHERE** wybrać **Zastosuj warunek WHERE**, a następnie w oknie **Łańcuch WHERE:** wpisać **year < 1944**. W **Opcje ►Oś x** odznaczyć **Pokaż linie siatki**. To samo zrobić dla osi *y*. W **Opcje ►Rozmiar wykresu** zmienić wymiary na szerokość 6 cali i wysokość 2 cale. Kliknąć **Uruchom**.

Zadanie 5. Zbiór danych **mortality** zawiera wskaźniki umieralności wskutek czerniaka złośliwego dla białych mężczyzn w okresie 1950–1969 dla poszczególnych stanów USA. Dodatkowo podano długość i szerokość środka każdego stanu. Narysować wykresy punktowe umieralności względem szerokości geograficznej oraz umieralności względem długości geograficznej. W każdym przypadku określić współczynnik korelacji. Zinterpretować rezultaty.

Zadanie 6. Zbiór danych **index** zawiera wskaźnik cen żywności i wskaźnik cen domów w Wielkiej Brytanii w latach 1971–1989. Poprzez narysowanie odpowiednich wykresów zbadać jak oba wskaźniki zmieniały się w czasie oraz jak te zmiany są ze sobą związane. Proszę poeksperymentować ze współczynnikiem kształtu obrazu.

Zadanie 7. Zestaw danych **expenditure** pochodzi z badania wykonanego w 1989 r. w Wlk. Brytanii. Dwie zmienne są średnimi tygodniowymi wydatkami (w funtach) na alkohol i tytoń ponoszonymi w budżecie przeciętnej rodziny.

(a) Narysować wykres punktowy, identyfikując regiony w odpowiedni sposób.

(b) Obliczyć współczynnik korelacji dla dwóch zmiennych.

(c) Czy któreś z regionów powinny być wyłączone z obliczeń korelacji i dlaczego? Ewentualnie powtórzyć obliczenia po ich usunięciu i zaobserwować spowodowane tym zmiany.

Zadanie 8. Zbiór danych **iqs** zawiera ilorazy inteligencji 30 amerykańskich małżeństw (dane z 2001 r.).

(a) Narysować wykres punktowy i zidentyfikować obserwacje mogące zniekształcić analizę.

(b) Określić korelację między ilorazami inetligencji z włączeniem oraz bez włączania obserwacji zidentyfikowanych w poprzednim punkcie.

(c) Dopasować linię prostą, aby przewidywać iloraz inteligencji kobiety na podstawie ilorazu inteligencji mężczyzny, oraz zobrazować tę prostą na wykresie punktowym.

(d) Zbadać reszty z dopasowanego modelu. Czy odpowiednie wykresy dają podstawy do obaw co do dopasowanego modelu?

(e) Jakie światło rzucają te dane na hipotezę, że ludzie poszukują partnera z podobnym ilorazem inteligencji?

Zadanie 9. Zbiór danych **galaxies** zawiera dane o prędkości i odległości 24 galaktyk (dane zebrano za pomocą kosmicznego teleskopu Hubble'a).

(a) Narysować wykres punktowy tych danych.

(b) Znaleźć odpowiedni model liniowy dla tych danych.

(c) Użyć tego modelu do oceny wieku wszechświata.

Zadanie 10. Zbiór danych **olympic1500** zawiera czasy (w sekundach) zwycięzców w biegu na 1500 m na igrzyskach olimpijskich od 1896 do 2004 roku.

(a) Narysować wykres punktowy. Czy występują obserwacje, które powinny być usunięte z dalszej analizy?

(b) Dopasować linię prostą do danych (po usinięciu ewentualnych obserwacji zidentyfikowanych w poprzednim punkcie).

(c) Czy model liniowy jest sensowny dla tych danych? Podać powody.

(d) Jeśli model liniowy jest nieadekwatny, dopasować model, który miałby być adekwatny i narysować go na tle zbioru danych.

(e) Użyć skonstruowanego modelu do budowy 95% przedziałów ufności dla przewidywanych czasów zwycięzców na igrzyskach w Pekinie (2008 r.) oraz w Londynie (2012). Sprawdzić, jak się one mają do wyników faktycznych zwycięzców na tych igrzyskach).