„Metody predykcji – konstruowanie modelu regresji z wielowymiarowych danych”. Przewodnik do ćwiczenia nr 2 dla studentów TPD w ramach przedmiotu „Zaawansowana eksploracja danych” (październik 2008)

Autorzy: *Jerzy Stefanowski*, *Robert Susmaga* Instytut Informatyki, Politechnika Poznańska.

Cel ćwiczenia: Celem jest nabycie umiejętności w konstruowaniu właściwego modelu regresji wielowymiarowej linowej oraz nieliniowej, ocena jego poprawności oraz wykorzystanie do predykcji wartości zmiennej objaśnianej. Ponadto należy zapoznać się z metodami selekcji zmiennych do modelu regresji. Nabyta wiedza może być przydatna w dalszych zajęciach, zwłaszcza Case 3.

Wstępna wiedza: Studenci przed rozpoczęciem zajęć powinni przypomnieć sobie wiedzę na temat regresji prostej (przedmiot Statystka i analiza danych na 2 roku studiów) oraz z treścią ostatnich wykładów z niniejszego przedmiotu (podstawowe slajdy dostępne na stronie WWW J.Stefanowskiego). Ponadto należy przypomnieć sobie zasady wykorzystywania funkcji statystycznych w Excel (przede wszystkim REGLINP, ROZKŁAD.?.ODW, KORELACJA oraz dodatku Analiza danych). W przypadku braku notatek z poprzedniego przedmiotu proszę skorzystać z książek nt. statystycznej analizy danych i zapoznać się z pomocą / „helpem” programu Statistica.

Umiejętności do nauczenia się w trakcie zajęć: W dalszej części ćwiczenia nauczyć się wykorzystania oprogramowania Statsoft Statistica (opcja Statistics / MultipleRegression oraz Wykresy (Graphs)). Ponadto dla regresji nieliniowej skorzystaj z opcji Zaawansowane model liniowe/nieliniowa (Advanced linear/nonlinear models) i następnie przejdź do nieliniowej estymacji.

Sugerowany przebieg ćwiczenia: Lista zadań do wykonania zostanie wskazana przez prowadzącego ćwiczenie. Niektóre z nich dotyczą podobnych zagadnień i mogą zostać wybrane lub pominięte w zależności od decyzji prowadzącego. Ponadto raporty z wykonania powinny być przedmiotem sprawozdania końcowego.

Raport końcowy: Powinien zawierać rozwiązania wybranych zadań, tzn. postać równań regresji, niezbędne wartości statystyk testowych (także wraz ze sformułowaniem odp. hipotez) oraz współczynników oceny równania regresji. W niektórych przypadkach konieczne będzie umieszczenie kopii wykresów – w szczególności dotyczy to analizy reszt i przypadków regresji nieliniowej. Ponadto, zapisz krótką słowną interpretację otrzymanych wyników i wnioski co do funkcji regresji, którą odkryłeś.

Lista zadań do wykonania:

Zadanie 1. Przypomnienie zasad badania zależności zmiennych liczbowych i budowania regresji prostej dwóch zmiennych.
Oporność R elementu elektrycznego (o charakterystyce liniowej) może być wyrażona jako:
R = U/I, gdzie:

- U jest napięciem na tym elemencie,
- I jest natężeniem prądu płynącego przez ten element.

Okazuje się, że ze względu na niedokładności pomiarów jednokrotne zmierzenie napięcia U_0 oraz natężenia I_0 i obliczenie R jako U_0/I_0 może być obarczone sporym błędem. Postawiono pytanie, czy można obliczyć lepsze oszacowanie oporności zakładając, że dysponujemy wielokrotnymi pomiarami napięć (y_i) i natężeń (x_i). Czy można sformułować wniosek dotyczący związku pomiędzy opornością a parametrami regresji liniowej?

W pierwszej kolejności spróbuj wykonać zadanie wykorzystując funkcje statystyczne Excel lub dodatek Analiza Danych.

x	y
0.004415	486.0979
0.008186	946.0276
0.007016	822.3864
0.006204	736.704
0.006729	839.794
0.005895	657.5608
0.004343	477.8808
0.002785	359.5351
0.001615	178.283
0.008906	1124.66
0.001196	131.3007
0.009098	1164.702
0.008356	1035.581
0.001646	213.7894
0.001392	168.3533
0.005566	639.9905
0.00354	407.7721
0.009376	1036.503
0.001485	168.5059
0.007694	1004.957

Zadanie 2. Ocena statystycznej jakości modelu regresji liniowej

W załączonym skoroszycie regresjaocena.xls wybierz drugi skoroszyt „Samochod". Opisuje one cenę pewnego modelu samochodu w zależności od jego rocznika (wieku). Dokonaj wykresu zmiennej *Cena* w zależności od *Wieku*. Na utworzonym wykresie dodaj linię trendu oraz równanie.

Stosując funkcję REGLINP znajdź estymaty wyrazu wolnego b_0 oraz współczynnika regresji b_1. Następnie odnajdź lub wygeneruj pełny raport zawierający: błędy standardowe parametrów, współczynniki determinacji, błąd standardowy reszt, regresyjne sumy kwadratów, resztowe sumy kwadratów oraz wartości statystyki F.

Przeprowadź ocenę jakości stworzonego modelu, interpretując powyższe wartości współczynników. Jeśli ich wartości są statystycznie istotne, dokonaj globalnej oceny istotności modelu z wykorzystaniem statystyki F (Przyjmij poziom istotności p=0,05).

Oblicz przedziały ufności dla parametrów modelu.

Jeśli jesteś przekonany do wiarygodności znalezionego modelu oblicz przewidywaną wartość y dla każdej historycznej realizacji *x* (wieku pojazdu). Oceń wielkości reszt.

Ew. wykonaj te ćwiczenia z wykorzystaniem oprogramowania Statsoft Statistica

Zadanie 3. Ocena statystycznej poprawności modelu – wersja z wykorzystaniem programu Statsoft Statistica

Dla próby 18 krajów oraz dodatkowo Japonii, USA i Chin przeprowadzono badania oceny procentowego tempa przyrostu PKB w 2000 roku (zmienna objaśnianego X) oraz procentowego tempa inflacji (zamienna objaśniana Y). Dane dla poszczególnych krajów zawarte są w poniższej tabeli (dane na podstawie książki A.Luszniewicz, T.Słaby: Statystyka z pakietem Statistica PL, Teoria i zastosowania).

	PKB%	Inflacja
1 Polska	4,5	7,5
2 Bulgaria	3,0	8,5
3 Chiny	7,0	2,5
4 Czechy	1,5	4,6
5 Estonia	4,0	4,7
6 Francja	2,7	1,1
7 Grecja	3,7	2
8 Hiszpania	2,9	2,3
9 Irlandia	6,8	2,6
10 Japonia	0	0,1
11 Litwa	2,5	4,3
12 Łotwa	2,5	3
13 Niemcy	2,3	1,4
14 Rosja	1,0	38
15 Rumunia	1,0	28
16 Słowacja	0	18
17 Ukraina	0	20
18 USA	2,7	2,6
19 Węgry	3,0	9,7
20 Włochy	2,1	2
21 W.Brytania	2,6	2,6

Przeprowadź analizę regresji i dokonaj weryfikacji statystycznej modelu. Najpierw wprowadź dane do oprogramowanie Statistica. Zdefiniuj plik w opcji „Nowe dane". Nazwij odpowiednio zmienne oraz przypadki. Warto wprowadzić nazwy przypadków.

W pierwszej kolejności zapoznaj się z opisem własności zmiennych poprzez wybór procedur w ramach Statystyki opisowe (skorzystaj w opcji więcej). Warto obliczyć współczynniki zmienności obu zmiennych – oceń czy charakteryzuje je wystarczająca zmienność. Wykonaj wykres zależności X od Y (Procedury Wykres, wybór 2D rozrzutu (scaterplots), opcje bez linii, albo z linią – wtedy warto zaznaczyć podopcje związane z obliczaniem regresji, przedziałów 0.95 ufności oraz współczynnika R2. Dokonaj optycznej analizy wykresu i zidentyfikuj linię regresji. Jeżeli uznajesz, że wykres może świadczyć o zależności liniowej, to powracając do Statystyk Podstawowych oblicz współczynnik korelacji oraz wykonaj test istotności współczynnika korelacji. Zinterpretuj wyniki. Przeprowadź analizę wnioskowania z linii funkcji regresji prostej. W menu procedury Regresja wieloraka. Wybierz odpowiednią zmienną zależną i niezależną. Nie włączaj opcji regresji krokowej, grzbietowaj itp. pozostań przy trybie podstawowym. Potwierdź OK. i przejdź do wyników regresji. Wyświetl podsumowanie regresji i raport ANOVA – przeprowadź analizę wyników i oceń statystyczną istotność modelu na poziomie lokalnym (współczynników) i globalnym (test F).

Jeżeli uznajesz, że model jest statystycznie wiarygodny, to przeprowadź predykcję wartości inflacji dla następujących możliwych wartości PKB 2,65%. Oszacuj przedział ufności wokół prognozowanej wartości.

Zadanie 4. Ocena statystycznej poprawności modelu – wersja 2

Podczas badania ruchu ulicznego przeprowadzono obserwacje długości drogi hamowania (y_i, [m]) w zależności od prędkości pojazdu (x_i, [km/h]). Zależność ta jest dość skomplikowana, ale – upraszczając problem – można uznać ją za liniową, tzn. mającą postać $y = b_1x + b_0$. Jakie wartości parametrów b_1 i b_0 najlepiej oddają tę zależność?

x	21.53	34.18	16.00	37.01	18.77	28.96	22.59	38.69	35.40	31.18	32.82	11.26	25.40	38.20	27.35	26.75
y	11.00	15.86	5.49	16.47	4.27	25.63	24.40	21.35	20.13	17.08	9.72	1.22	12.20	28.36	15.25	12.70
x	29.01	16.12	28.60	17.79	30.17	30.99	12.87	25.74	30.57	20.52	6.40	24.11	20.28	22.35	14.48	37.98
y	12.81	10.37	23.18	5.18	14.03	10.90	4.88	9.77	20.74	14.30	0.61	6.09	7.93	18.30	3.05	28.06
x	16.09	17.70	28.26	6.30	32.68	32.18	39.60	40.23	23.99	20.92	22.53	11.20	27.35	19.40	24.10	19.31
y	7.93	8.54	17.08	3.00	19.52	14.64	36.60	25.92	16.47	10.37	7.93	6.70	9.76	8.54	7.70	6.10

Dokonaj pełnej analizy poprawności stworzonego modelu. Wykonaj wykres rozrzutu X, Y , aby zorientować się dokładniej w przebiegu funkcji. Stwórz pełen raport wskaźników oraz testów charakteryzujących otrzymany model regresji. Zinterpretuj m.in.: współczynniki R^2, współczynniki błędu, wyniki testów istotności współczynników równania regresji, rozrzut reszt.
Jeżeli weryfikacja modelu przebiegła poprawnie i wyniki mogą być zaakceptowane, to dokonaj predykcji wartości y na podstawie modelu dla $x = 11.20$, $x = 24.91$ oraz $x = 40.00$.

Zadanie 5. Regresji liniowa dla rzeczywistych danych

W załączonym pliku cereals.dat znajduja zbiór danych o charakterystyce płatków śniadaniowych dostępnych w przeszłości na rynku amerykańskim (patrz źródło Data and Story Library http://lib.stat.cmu.edu/DASL). Zbiór danych zawiera informacje o wartościach odżywczych 77 rodzajów płatków śniadaniowych i zawiera następujące atrybuty:

Nazwa płatków (cereal names).

Producent płatków (manuf).

Typ płatków (type) – nominalny atrybut (do jedzenia na ciepło – hot ; lub do jedzenia na zimno – cold).

Kalorie (calories) w porcji.

Białka (protein) w gramach.

Tłuszcz (fat) w gramach.

Sód (sodium) w miligramach.

Błonnik (fiber) w gramach.

Węglowodany (carbohydrates) w gramach.

Cukry (sugar) w gramach.

Potas (potass) w miligramach.

Witaminy (vitamins) – procent zalecanego dziennego spożycia witamin.

Waga porcji (weight)

Liczba łyżek (cups) w porcji.

Położenie półki (szelf): 1dolna, 2 środkowa, 3 górna

Wartość odżywcza (rating) – oszacowana przez Consumer reports.

W pierwszym etapie wykonać analizę tylko regresji prostej dla dwóch zmiennych, tj.

Zmienna zależna – wartość odżywcza.

Zmienna niezależna - cukry

Podobnie jak w poprzednich ćwiczeniach wykonać najpierw wykres rozrzutu i ocenić kształt potencjalnej funkcji regresji. Jeżeli uznajesz, że wykres może świadczyć o zależności liniowej, to przechodząc do procedur Regresja wieloraka przeprowadź analizę wnioskowania z linii funkcji regresji prostej. Dodatkowo dokonaj analizy reszt – odpowiednie opcje w dialogu. Wykonaj wykres normalnego kwantylowego rozkładu reszt – oceń czy jest spełnione założenia o normalności rozkładu reszt. Następnie wykonaj i przeanalizuj kształt wykresu rozrzutu reszt wobec wartości przewidywanej. Jeżeli uznasz, że są obserwacje oddalone (ang. outliers) zidentyfikuj je na podstawie analizy wartości standaryzowanych reszt – inna z opcji w dialogu związanym z analizą reszt.

Podsumuj krótko wnioski z powyższych analiz i oceń jakość zbudowanego modelu.

Zadanie będzie kontynuowane dla wielu zmiennych.

Zadanie 6. Regresji liniowa wielu zmiennych – dobór zmiennych

W załączonym skoroszyt *regresjaocena.xls* wybierz trzeci arkusz „Dochod". Zawierają one informacje o dochodach pewnej firmy w ostatnich latach w odniesieniu do informacji o inwestycjach (dotacje) oraz średnich realnych zarobkach oraz kolejnych latach. Zbuduj model liniowej regresji wielowymiarowej oraz przeprowadź analizę istotności zmiennych – starając się przeprowadzić selekcję zmiennych (np. posługując się wartością dopasowanego współczynnika R-kwadratu).

Zadanie 7. Regresji liniowa wielu zmiennych – wersja regresji krokowej

W pliku *lekiczas.txt* znajdują się dane związane z oceną czasu pobytu pacjenta w klinice w trakcie leczenia w zależności od dawek czterech różnych specyfików. Przy pomocy regresji krokowej postępującej (ang. forward) i wstecznej (ang. backward) dokonaj selekcji zmiennych do modelu regresji liniowej, który jak najlepiej przewiduje czas pobytu chorego w szpitalu. Skorzystaj z oprogramowania Statistica procedury Regresja Wieloraka, wybierz opcje Krokowa zamiast trybu standardowego.

Zadanie 8. Budowa i ocena statystycznej poprawności modelu regresji wielowymiarowej

Dane w pliku *zf-examples-md1.xls* = Plik zawiera trzy zestawy danych:

- zestaw X,Y – zawierający zmienne X i Y
- zestaw X,Y,Z – zawierający zmienne X, Y i Z
- zestaw W,X,Y,Z – zawierający zmienne W, X, Y i Z

Sugerowany przebieg analizy:

Zestaw X,Y – wykorzystując mechanizm regresji jednowymiarowej znaleźć zależność pomiędzy zmiennymi X i Y oraz ocenić istotność modelu liniowego.

Zestaw X,Y,Z – wykorzystując mechanizm regresji jednowymiarowej znaleźć zależności pomiędzy parami zmiennych X i Y, X i Z oraz Y i Z. Można też zbudować macierz współczynników korelacji. Czy można stwierdzić, która ze zmiennych jest zmienną objaśnianą – zależną, a które zmienne są niezależne (objaśniające). Zbuduj ostatecznie model liniowej regresji wielowymiarowej zależności jednej zmiennej od dwóch pozostałych.

Zestaw W,X,Y,Z – zawierający zmienne W, X, Y i Z. Ustal, która ze zmiennych jest zmienną objaśnianą – zależną, a które zmienne są niezależne (objaśniające). Zbuduj ostatecznie model regresji wielowymiarowej

Zadanie 9. Budowa modelu regresji wielowymiarowej z selekcją zmiennych

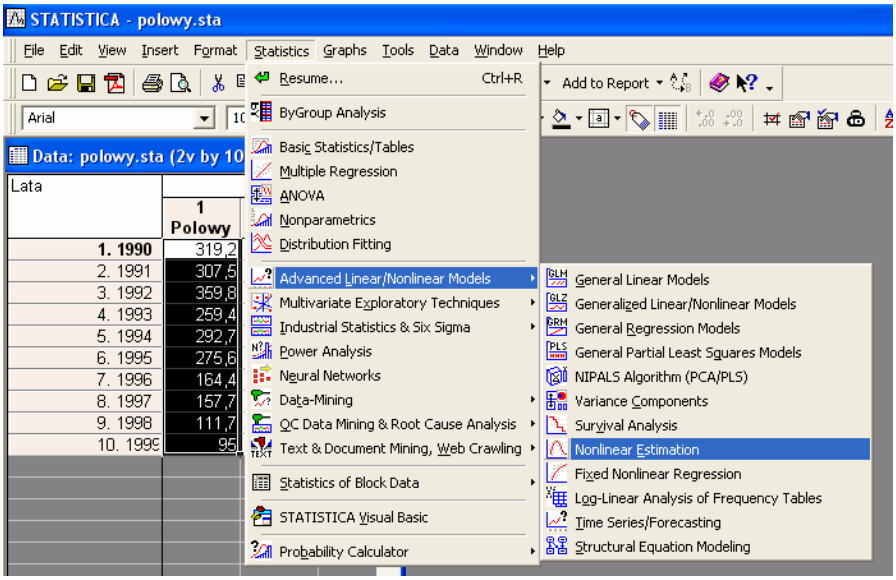
Powróćmy do zbioru *cereals.dat* w celu zbudowania modelu regresji wielowymiarowej. Chcemy jednak zbudować jak najlepszy model oszacowania wartości odżywczych płatków niezawierający jednak nieistotnych zmiennych. Przypomnijmy definicje atrybutów (patrz zadanie 5) i ustalmy, które z nich są zdefiniowane na skala jakościowych – te można pominąć. Zbuduj najpierw globalny model z wykorzystaniem wszystkich zmiennych i dokonaj oceny jego istotności statystycznej. Następne zastosuj różne metody selekcji zmiennych (patrz wykład – ocena korelacji, metoda regresji krokowej tzn. dołączania pojedynczych zmiennych – skorzystaj z właściwych opcji w oknie regresja wieloraka). Wybierz najlepszy model zawierający jak najmniej sensowych interpretacyjnie zmiennych. W trakcie poszukiwań możesz także sprawdzić czy występują obserwacje oddalone (jeśli uznasz że wpływają na wyniki, to usuń je i przeprowadź analizę dla zmniejszonego zbioru produktów-obserwacji // możesz skorzystać także z analizy reszt).

Zadanie 10. Budowa modelu regresji nieliniowej (mniejszy zbiór danych)

W pewnej książce przedstawiono dane nt. badania potencjału polskiego rybołówstwa dalekomorskiego w latach 1990-1999, gdzie zestawiono między innymi rozmiary połowów (w tys. ton) i liczbę statków połowowych. Dane są przedstawione poniżej:

Lata	Polowy	Statki
1. 1990	319,2	77
2. 1991	307,5	65
3. 1992	359,8	53
4. 1993	259,4	45
5. 1994	292,7	42
6. 1995	275,6	36
7. 1996	164,4	33
8. 1997	157,7	33
9. 1998	111,7	32
10. 1999	95	31

Załóżmy, że chcesz zbudować model regresji dla tych danych. Najpierw zrób wykres rozrzutu XY, aby ocenić kształt potencjalnej zależności. Jeżeli uznasz, że jest on silnie nieliniowy postaraj się dobrać najlepszy z możliwych kształtów funkcji nieliniowych (hiperboliczna, logarytmiczna, kwadratowa lub potęgowa). Wykorzystaj wiedzę o wartości współczynnika R2 i ew. innych metod oceny dopasowania empirycznej funkcji do rozrzutu punktów w danych. Uwaga: zamiast dotychczasowej opcji Regresja wieloraka, zalec się przejść do innej grupy procedur związanych z estymacją nieliniową. To znaczy w ramach opcji rozwijanej menu głównego Statystyki wybierz „Zawansowane model liniowe/nieliniowe" (Advanced linear/nonlinear models) i następnie przejdź do nieliniowej estymacji (patrz rysunek poniżej). W tej wersji funkcje transformująca użytkownik definiuje sam w specjalnym dialogu z wykorzystaniem typowych operatorów i nazw zmiennych (wybierz wtedy wersję funkcja regresji określona przez użytkownika i w kolejnym oknie „funkcja estymowana" zdefiniuj postać funkcji).



Innym sposobem postępowania jest wykorzystanie niżej opcji Ustalona funkcja nieliniowa (ang. fixed nonlinear regression), gdzie po wybraniu stosownych zmiennych, możesz w kolejnym dialogu

„nieliniowe składniki regresji” możesz wybrać rozważane operatory transformacji. W obu przypadkach późniejsze okna dialogowe do obsługi wyników i analizy reszt przypominają rozwiązania z opcji regresji wielokrotnej.

Zadanie 11. Budowa modelu regresji nieliniowej

Dane w pliku *non-lin-data.xls* -> dane eksperymentalne dotyczą pewnego zjawiska nieliniowego (X – zmienna niezależna, Y – zmienna zależna). Wykorzystać technikę znajdowania współczynników wielowymiarowej regresji liniowej do znalezienia przybliżonej zależności pomiędzy zmiennymi X i Y (Y=f(X)).

- jaka postać funkcji nieliniowej najlepiej przybliża zależność pomiędzy X i Y?
- jak wykorzystać informacje o postaci funkcji nieliniowej do wyznaczenia współczynników modelu zależności pomiędzy zmiennymi?

Oceń jakość otrzymanego modelu.