Analiza zależności cech ilościowych— regresja liniowa (Wykład 13)

dr Mariusz Grządziel

semestr letni 2012

Przykład wprowadzający
W zbiorze danych `homedata` (z pakietu R-owskiego UsingR) można znaleźć ceny 6841 domów Maplewood (New Jersey) z lat: 1970 (zmienna `y1970`) i 2000 (zmienna `y2000`). Interesuje nas zależność pomiędzy cenami domów: z roku 1970 i 2000.

Dane dotyczące cen domów w Maplewood
Dane (w dolarach) dotyczące pierwszych 12 domów z tego zbioru danych:

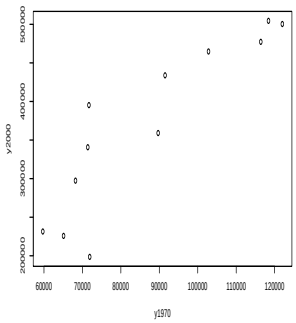
```
> homedata[1:12,]
    y1970  y2000
1   89700 359100
2  118400 504500
3  116400 477300
4  122000 500400
5   91500 433900
6  102800 464800
7   71700 395300
8   71400 340700
9   68200 297400
10  71900 198600
11  65100 225800
12  59700 231500
```

Te skrócone dane zostaną zapisane do zbioru `h1`.

Wykres rozproszenia

Dopasowanie prostej do chmury danych
W oparciu o wykres rozproszenia można próbować ocenić istnienie i charakter zależności zmiennej y2000 i zmiennej y1970.
Problem: w jaki sposób dobrać prostą (o równaniu $y = b_0 + b_1 x$) tak, aby najlepiej pasowała do „chmury danych" przedstawionej na wykresie rozproszenia.

Liniowa zależność pomiędzy dwiema zmiennymi, prosta regresji
Rozważmy przypadek ogólny
Dane: próba dwucechowa $(x_1, y_1), \ldots, (x_n, y_n)$
W naszym przykładzie: $n = 12$, x-y odpowiadają cenom z roku 1970, y-eki cenom z roku 2000.



Rysunek 1: Wykresy rozproszenia dla danych dotyczących `h1`

Prosta MNK
Chcemy „dopasować" prostą $y = b_0 + b_1 x$ do naszych danych (chmury punktów). Dla danego x_i wartość $\hat{y}_i = b_0 + b_1 x_i$ można interpretować jako wartość y przewidywaną na podstawie rozpatrywanej prostej dla wartości zmiennej objaśniającej równej x_i. Błąd oszacowania, czyli tzw. wartość resztowa lub rezyduum wynosi $y_i - \hat{y}_i$.
Chcemy znaleźć prostą $y = b_0 + b_1 x$, dla której kwadratów rezyduów

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 \tag{1}$$

jest minimalna.
Współczynnik b_1 nazywamy *współczynnikiem kierunkowym* a b_0 *wyrazem wolnym*.

Liniowa zależność pomiędzy dwiema zmiennymi, prosta regresji

Definicja 1. *Prostą regresji opartą na metodzie najmniejszych kwadratów nazywamy prostą, dla której wartość sumy $S(b_0, b_1)$ w (1) traktowanej jako funkcja wszystkih możliwych wartości współczynnika kierunkowego i wyrazu wolnego, jest minimalna.*

Nazwy: prosta regresji, prosta MNK.

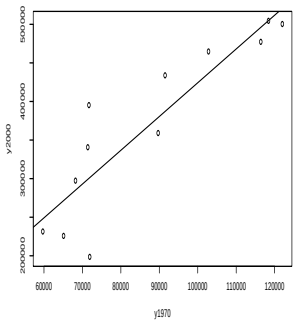
Liniowa zależność pomiędzy dwiema zmiennymi, prosta regresji
Stosując podstawowe techniki znajdowania mimimum funkcji dwóch zmiennych i proste przekształcenia algebraiczne otrzymujemy:

$$b_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

gdzie $\bar{x}$ i $\bar{y}$ oznaczją średnie dla $x_1, \ldots, x_n$ i $y_1, \ldots, y_n$ oraz

$$b_0 = \frac{1}{n} \big(\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i \big).$$

Wartość $y = b_0 + b_1 x$ nazywamy wartością przewidywaną zmiennej objaśnianej na podstawie prostej MNK dla wartości zmiennej objaśniającej równej x.



Rysunek 2: Wykres rozproszenia+prosta MNK dla danych `h1`

Dane h1:Wykres rozproszenia+prosta MNK

Rozkład całkowitej zmienności zmiennej objaśnianej
Oznaczmy

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2}$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{3}$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4}$$

gdzie SST:całkowita suma kwadratów (Total Sum of Squares), SSR regresyjna suma kwadratów, SSE suma kwadratów błędów.

Twierdzenie 1. *Prawdziwa jest równość*

$$SST = SSR + SSE.$$

Dowód można znaleźć w książce Koronackiego i Mielniczuka, str. 270.
Współczynnik SSR/SST zwany współczynnikiem determinacji (oznaczanym R^2) określa stopień, w jakim zależność liniowa między zmienną objaśnianą a objaśniającą tłumaczy zmienność wykresu rozproszenia.

Współczynnik korelacji

Definicja 2. *Współczynnikiem korelacji próbkowej dla próbki dwuwymiarowej* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ *nazywamy zmienną liczbę*

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

gdzie $\bar{x}$ i s_x oznaczają średnią i odchylenie standardowe próby $x_1, x_2, \ldots, x_n$ i podobnie $\bar{y}$ i s_y oznaczają średnią i odchylenie standardowe próby $y_1, y_2, \ldots, y_n$.
Uwaga. W powyższej definicji przez n-elementową próbę dwuwymiarową rozumiemy ciąg n par liczb (a nie zmiennych losowych).

Współczynnik determinacji a współczynnik korelacji
Można udowodnić, że współczynnik korelacji przyjmuje wartości z przedziału $[-1, 1]$. Wartości współczynnika korelacji bliskie 1 lub -1 wskazują na istotną zależność liniową pomiędzy zmiennymi.

Twierdzenie 2. *Zachodzi równość*

$$r^2 = SSR/SST.$$

Przykład z cenami domów w Maplewood— c.d.
Wykonując obliczenia dla danych $h1$ otrzymujemy

$$R^2 = RRS/RRT = 0{,}7944 \quad r = 0{,}891287.$$

Problem: powyższe wartości współczynników: determinacji i korelacji wskazują na istotną zależność pomiędzy cenami domów w latach 1970 i 2000?

Statystyczny model zależności liniowej
Rozważmy *model regresji liniowej:*

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n, \tag{5}$$

gdzie β_1 i β_2 są pewnymi stałymi a $\epsilon_1, \ldots, \epsilon_n$ są niezależnymi zmiennymi losowymi o rozkładzie $N(0, \sigma)$. Przyjmujemy ponadto, że wartości x_i nie są sobie równe (nie są równe jednej liczbie).
Mamy tu n zmiennych losowych: $Y_1, Y_2, \ldots, Y_n$.
Dla zmiennej Y_i wartość oczekiwana jest równa:

$$\mu_{Y_i} = E(\beta_0 + \beta_1 x_i) + E(\epsilon_i) = \beta_0 + \beta_1 x_i, \qquad i = 1, \ldots, n.$$

Sens: Wartość objaśniana jest równa funkcji liniowej zmiennej objaśnianej plus pewien błąd losowy.

Wnioskowanie statystyczne w modelu zależności liniowej
Możemy być zainteresowani:

- estymacją parametrów β_0, β_1,
- estymacją wariancji σ^2,
- estymacją przedziałową ww. parametrów lub ich funkcji;
- testowaniem hipotez dotyczących parametrów.

Weryfikacja hipotezy dotyczącej istotności β_1
Jesteśmy zainteresowani weryfikacją

$$H_0 : \beta_1 = 0 \text{ przeciw } H_1 : \beta_1 \neq 0.$$

Stosowną statystyką testową okazuje się

$$t = \frac{b_1}{\mathcal{S}} \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

gdzie

$$b_1 = \frac{\sum_{i=1}^{n} x_i (Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

$$\mathcal{S}^2 = \frac{\sum_{i=1}^{n} (Y_i - b_1 x_i - b_0)^2}{n-2} \quad \text{z } b_0 = \bar{Y} - b_1 \bar{x}.$$

Weryfikacja hipotezy dotyczącej istotności β_1-c.d.
Można pokazać, że

$$t \sim t_{n-2} \quad \text{(t ma rozkład t-Studenta z } n-2 \text{ st. swobody).}$$

Obszarem krytycznym dla poziomu istotności α jest:

$$(-\infty, -t_{1-\alpha/2,n-2}] \cup [t_{1-\alpha/2,n-2}, \infty)$$

Hipotezę H_0 przeciwko H_1 można również testować opierając się na statystyce

$$F = \frac{SSR}{SSE/(n-2)}.$$

W powyższym wzorze SSR i SSE obliczamy kładąc Y_i zamiast y_i w odpowiednich wzorach (por. str. 10). Statystyka F ma rozkład $F_{1,n-2}$.

Obliczenia w środowisku R

```
> reg<-lm(y2000~y1970,data=h1)
> summary(reg)

Call:
lm(formula = y2000 ~ y1970, data = h1)

Residuals:
    Min      1Q  Median      3Q     Max
-102658  -20099   -8331   31617   94918

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.348e+04  6.341e+04  -0.213    0.836
y1970        4.377e+00  7.042e-01   6.216 9.94e-05 ***
--
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 52860 on 10 degrees of freedom
Multiple R-Squared: 0.7944,     Adjusted R-squared: 0.7738
F-statistic: 38.64 on 1 and 10 DF,  p-value: 9.938e-05
```

p-wartość odpowiadająca weryfikacji

$$H_0 : \beta_1 = 0 \text{ przeciw } H_1 : \beta_1 \neq 0.$$

jest równa $9.938e-05 = 0{,}0000938$— są podstawy do odrzucenia H_0 nawet dla poziomu istotności $\alpha = 0{,}0001$.

Problem prognozy
Jesteśmy zainteresowani ceną, za którą można byłoby sprzedać pewien dom w Maplewood, który w 1970 r. kosztował 100000; (informacje dotyczące ceny tego domu w 2000 r. nie znajdują się w naszym zbiorze danych).

Problem prognozy— rozważania dla ogólnego przypadku
W modelu zależności liniowej jesteśmy zainteresowani oszacowaniem wartości oczekiwanej:
Prognozą zmiennej zależnej Y dla ustalonej wartości x_0 nazywamy zmienną losową:

$$Y(x_0) = \beta_0 + \beta_1 x_0.$$

Sensowne oszacowanie wartości oczekiwanej $Y(x_0)$:

$$\bar{Y}(x_0) = b_0 + b_1 x_0.$$

Przedział ufności na poziomie ufności $1-\alpha$ dla wartości oczekiwanej $Y(x_0)$:

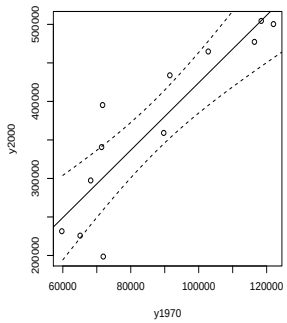
$$\bar{Y}(x_0) \pm t_{1-\alpha/2,n-2} SE_{\bar{Y}(x_0)}$$

gdzie $SE_{\bar{Y}(x_0)} = \mathcal{S}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

Problem prognozy— przykład
Chcemy znaleźć wartość oczekiwaną $Y(100000)$ (dla danych `h1`) i 95-procentowy przedział ufności dla $Y(100000)$.
Poleceniem systemu R, przy pomocy którego można wykonać odpowiednie obliczenia (wyżej opisane) jest predict; dla naszych konkretnych danych należy je zastosować w następujący sposób:

```
> predict(reg,data.frame(y1970=100000),
interval="confidence",level=0.95)
          fit      lwr      upr
[1,] 424264.2 384932.2 463596.1
```

a więc 95-procentowym przedziałem ufności dla $Y(100000)$ jest $(384932{,}2; 463596{,}1)$.



Rysunek 3: Prosta MNK i 95-procentowe krzywe ufności

Krzywe ufności
Z połączenia końców przedziałów ufności dla $Y(\tilde{x}_i)$ dla odpowiednio dobranego ciągu (skończonego) $\tilde{x}_1, \ldots, \tilde{x}_m$, „wypełniającego" przedział zmienności zmiennej niezależnej $x_{min}, \ldots, x_{max}$ otrzymujemy tzw. krzywe ufności. Dodanie tych krzywych do wykresu rozproszenia z zaznaczoną prostą regresji pozwala na odczytanie z wykresu końców przedziału ufności (lub ich przybliżonej wartości) dla $Y(\tilde{x}_0)$, gdzie $x_0 \in [x_{min}, x_{max}]$

Krzywe ufności—c.d.

Weryfikacja poprawności modelu
W naszych rozważaniach założyliśmy, że dane z `h1`.
Założenie (o adekwatności modelu) można weryfikować analizując wartości resztowe (rezydua)— por. Koronacki i Mielniczuk (2001), par. 4.2.5 (str. 284–291).

Polecana literatura
R. Kala, Statystyka dla przyrodników, wyd. Akademii Rolniczej w Poznaniu 2002, rodz. B1-B3.
J. Koronacki i J. Mielniczuk, Statystyka dla studentów kierunków technicznych i przyrodniczych, WNT 2001, rozdz. 4.1-2.