Elementy statystyki

STA - Wykład 3

dr hab. Waldemar Wołyński
Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza

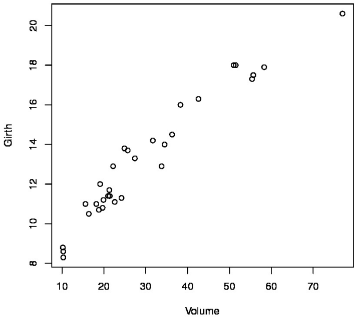
Prosta regresja liniowa

Model: prosta regresja liniowa

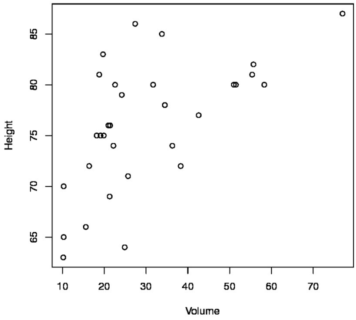
Modelujemy wyniki doświadczenia w którym dokonujemy n-niezależnych obserwacji zmiennej (zależnej) Y (typu ilościowego ciągłego) oraz niezależnej zmiennej X (typu ilościowego), związanych zależnością liniową na losowo wybranych z populacji jednostkach eksperymentalnych.

Przykład Staramy się oszacować objętość tarcicy (zmienna zależna) jaką można pozyskać z drzew czarnej wiśni. Objętość tarcicy zależy od wielkości drzewa opisanej przez jego średnicę i wysokość (zmienne niezależne). Dane dla 31 drzew zawarte są w ramce "trees" dostępnej w programie **R**. Budowę modelu rozpoczynamy od wykonania diagramów korelacyjnych.

Model: prosta regresja liniowa



objętość – średnica



objętość – wysokość

Jako zmienną niezależną przyjmujemy średnicę (dodatkowo przyjmujemy liniową zależność pomiędzy objętością a średnicą).

Model: prosta regresja liniowa

Wybieramy model prostej regresji liniowej postaci:

$$Y_i = a + bx_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

gdzie

Y_i – i-ta obserwacja zmiennej zależnej Y,
x_i – i-ta wartość zmiennej niezależnej X,
a, b – parametry liniowej funkcji regresji,
ε_i – błędy (reszty).

Uwaga:

W rozważanym przykładzie zmienna niezależna to średnica drzewa, zmienna zależna to objętość drzewa.

Założenia

O błędach zakładamy, że:

- są niezależne (dokładnie: są niezależnymi zmiennymi losowymi),
- mają wartość oczekiwaną równą zero (nie ma błędu systematycznego), tzn.

$$E(\varepsilon_i) = 0, \quad i = 1, \ldots, n,$$

- mają jednakową, stałą i niezerową wariancję, tzn.

$$Var(\varepsilon_i) = \sigma^2, \quad i = 1, \ldots, n.$$

Uwaga:

Model prostej regresji liniowej ma trzy parametry: a, b i σ^2.

Estymacja parametrów a i b liniowej funkcji regresji

Do estymacji parametrów a i b używamy metody **najmniejszych kwadratów**, polegającej na minimalizacji sumy kwadratów błędów, tzn.

$$S(a,b) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

Fakt

Estymatorami najmniejszych kwadratów (ENK) parametrów a i b liniowej funkcji regresji są statystyki:

$$\hat{a} = \bar{Y} - \hat{b}\bar{x}, \quad \hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Estymatory parametrów modelu

Twierdzenie

W modelu prostej regresji liniowej, statystyki $\hat{a}$ i $\hat{b}$ są nieobciążonymi estymatorami parametrów a i b.

Ponadto, statystyka

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}x_i)^2$$

jest nieobciążonym estymatorem parametru σ^2.

Estymatory parametrów modelu

Twierdzenie

Przy dodatkowym założeniu normalności rozkładu błędów, w modelu prostej regresji liniowej, statystyki

$$\hat{a} \quad \text{i} \quad S^2 \quad \text{oraz} \quad \hat{b} \quad \text{i} \quad S^2$$

są niezależnymi zmiennymi losowymi.

Ponadto

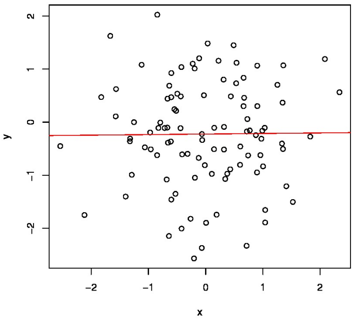
$$\hat{a} \sim N\left(a, \sigma^2\left(\frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right),$$

$$\hat{b} \sim N\left(b, \sigma^2\left(\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right)$$

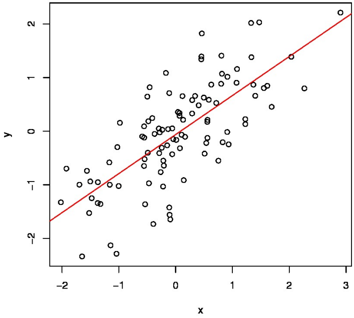
oraz

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2).$$

Wpływ zmiennej niezależnej X na zmienną zależną Y



Brak istotnego wpływu
$b = 0$



Istotny wpływ
$b \neq 0$

Dopasowanie modelu

W modelu prostej regresji liniowej prawdziwa jest następująca zależność:

$$SST = SSR + SSE$$

gdzie

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2, \quad SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_i)^2,$$

$$SSE = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2, \quad \hat{Y}_i = \hat{a} + \hat{b}x_i.$$

Liczbową miarą dopasowania prostej regresji do danych empirycznych jest **współczynnik determinacji** (podawany w %)

$$R^2 = 1 - \frac{SSE}{SST}, \quad 0 \leq R^2 \leq 1.$$

Prognozowanie (predykcja)

Model prostej regresji liniowej wykorzystujemy często do wyznaczania prognozy wartości zmiennej zależnej, przy ustalonej wartości zmiennej niezależnej. Niech x_p oznacza wartość zmiennej niezależnej X dla której uzyskać chcemy prognozę zmiennej zależnej Y równą Y_p.
Przyjmujemy:

$$\hat{Y}_p = \hat{a} + \hat{b}x_p.$$

Jako ocenę jakości prognozy przyjmujemy oszacowanie odchylenia standardowego prognozy (średni błąd prognozy) postaci:

$$S_p = S\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Wielokrotna regresja liniowa

Model: wielokrotna (wieloraka) regresja liniowa

Model ten jest rozwinięciem modelu prostej regresji liniowej na przypadek wielu zmiennych niezależnych.

$$Y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_m x_{im} + \varepsilon_i, \quad i = 1, \ldots, n,$$

gdzie

Y_i – i-ta obserwacja zmiennej zależnej Y,
$x_{i1}, x_{i2}, \ldots, x_{im}$ – i-te wartości zmiennych niezależnych $X_1, X_2, \ldots, X_m$,
$a_0, a_1, \ldots, a_m$ – parametry liniowej funkcji regresji,
ε_i – błędy (reszty).

Uwagi:

Założenia dotyczące błędów są identyczne jak w modelu prostej regresji liniowej.
Model wielokrotnej regresji liniowej ma $m + 2$ parametry: $a_0, a_1, \ldots, a_m$ i σ^2.

Zapis macierzowy modelu

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nm} \end{bmatrix}, \quad \boldsymbol{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix}. \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Model wielokrotnej regresji liniowej: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{\varepsilon}$.

Uwaga: W modelu wielokrotnej regresji liniowej zakładamy, że $n > m$ oraz $\mathrm{rząd}(\boldsymbol{X}) = m + 1$.

Estymatory parametrów modelu

Twierdzenie

W modelu wielokrotnej regresji liniowej, statystyka

$$\hat{\boldsymbol{a}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

jest nieobciążonym estymatorem parametru $\hat{\boldsymbol{a}}$.

Ponadto, statystyka

$$\hat{\sigma^2} = S^2 = \frac{1}{n-m-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

gdzie

$$\hat{Y}_i = \hat{a}_0 + \hat{a}_1 x_{i1} + \cdots + \hat{a}_m x_{im}, \quad i = 1, \ldots, n,$$

jest nieobciążonym estymatorem parametru σ^2.

Dopasowanie modelu

W modelu wielokrotnej regresji liniowej (analogicznie jak w modelu prostej regresji liniowej) prawdziwa jest następująca zależność:

$$SST = SSR + SSE.$$

Liczbową miarą dopasowania funkcji regresji (hiperpłaszczyzny regresji) do danych empirycznych jest **poprawiony współczynnik determinacji** (podawany w %)

$$R^2_{pop} = 1 - \frac{SSE/(n-m-1)}{SST/(n-1)}, \quad 0 \leq R^2_{pop} \leq 1.$$

Prognozowanie (predykcja)

Niech $\boldsymbol{X}_p = [1, x_1^p, \ldots, x_m^p]'$ oznacza wektor wartości zmiennych objaśniających $X_1, X_2, \ldots X_m$ dla którego uzyskać chcemy prognozę zmiennej objaśnianej Y.

Przyjmujemy:

$$\hat{Y}_p = \boldsymbol{X}_p' \hat{\boldsymbol{a}}.$$

Jako ocenę jakości prognozy przyjmujemy oszacowanie odchylenia standardowego prognozy (średni błąd prognozy) postaci:

$$S_p = S\sqrt{\boldsymbol{X}_p'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_p}.$$

R

Funkcje związane z modelem regresji liniowej (prostej i wielokrotnej):

lm – funkcja podstawowa,

summary – wartości estymatorów parametrów modelu regresji, wartość współczynnika determinacji, itp.

predict – wartości prognozowane,

abline – wykres prostej regresji.