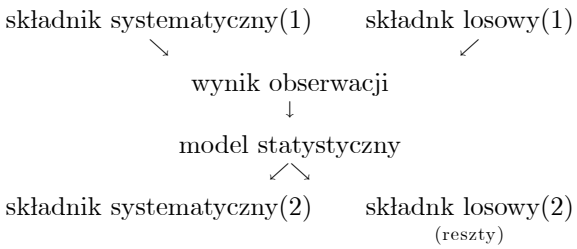
1 Regresja liniowa cz. I

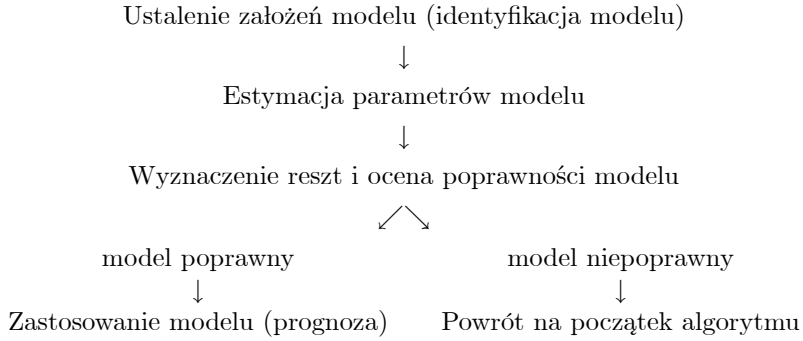
1.1 Model statystyczny

Model statystyczny to zbiór założeń. Wprowadzamy model, który możliwie najlepiej opisuje ineresujący nas fragment rzeczywistość. Błędy modelu wynikają z nieuwzględnienia wszystkich czynników, błędnego uwzględenia czynników oraz z losowego zaburzenia. Model statystyczny ma opisywać składnik systematyczny, to co pozostaje to losowe reszty.



Zakładamy, że składnik losowy (1)(zaburzenie losowe, błąd losowy) jest wynikiem działania mało istotnych zdarzeń. Czego spodziewamy się od reszt? Powinny być niezależne o średniej 0, i stałej wariancji (równomiernie rozłożona chmurka punktów), często zakłada się, że mają rozkład normalny (pomocne założenie w technikach obliczeniowych).

Budowa modelu statystycznego:



1.2 Model regresji liniowej

I ETAP: Ustalenie założeń modelu

Ze względu na przyjęte założenie o liniowej zależności między zmiennymi nasz model jest postaci:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Y jest zmienną zależną (inaczej zmienną objaśnianą), której wartości chcemy wyjaśnić lub przewidzieć

X jest zmienną niezależną (inaczej zmienną objaśniającą) nazywaną też predyktorem, zakładamy, że zmienna ta nie jest zdegenerowana do stałej. W przeciwnym przypadku problemu regresji nie byłoby sensu rozważać.

Uwaga 1 *Umawiamy się, że wartości X są ustalone (brak losowości).*

ε jest błędem losowym (inaczej zakłóceniem, szumem), jedynym źródłem losowości

β_0 to wyraz wolny będący punktem przecięcia linii $Y = \beta_0 + \beta_1 X$ z osią rzędnych.

β_1 jest współczynnikiem kierunkowym, czyli tangensem kąta pod którym linia $Y = \beta_0 + \beta_1 X$ nachylona jest do osi odciętych.

$$\begin{aligned} Y &= \underbrace{\beta_0 + \beta_1 X}_{\text{składnik systematyczny}} + \underbrace{\varepsilon}_{\text{składnik losowy}} \\ y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, ..., n \end{aligned}$$

Dodatkowe założenia modelu:

ε_i są nieskorelowane, o tym samym rozkładzie

$$E(\varepsilon) = 0$$
$$Var(\varepsilon) = \sigma^2$$

Jeśli dodatkowo

$$\varepsilon \sim N\left(0, \sigma^2\right)$$

to mówimy o modelu normalnej regresji liniowej.

Przykład 1 *RYSUNEK!!!*

Uwaga 2 *W modelu regresji obie zmienne są losowe, a w regresji tylko Y jest losowa.*

Własności 1 *W modelu normalnej regresji liniowej: $\varepsilon_i \perp \varepsilon_j$, $i \neq j$*

Uwaga 3 ***Własności 2*** *Błędy (ε) są jednakowo rozproszone wokół linii regresji tzn. ich rozkład jest identyczny (w tym identyczna średnia i odchylenie)*

Własności 3

1. $E(Y) = \beta_0 + \beta_1 X$, $(E(Y|X) = \beta_0 + \beta_1 X)$
2. $Var(Y) = \sigma^2$, $(Var(Y|X) = \sigma^2)$
3. $Cov(\varepsilon_j, y_i) = \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j \end{cases}$

II Etap: Estymacja parametrów modelu (obecny wykład skoncentruje się wokół tego etapu)

Estymatory parametrów β_0 i β_1 oznaczamy odpowiednio $\widehat{\beta}_0$ i $\widehat{\beta}_1$. Na podstawie n-elementowej próby (x_i, y_i), $i = 1, ..., n$ oszacowany model regresji jest postaci:

$$Y = \widehat{\beta}_0 + \widehat{\beta}_1 X + e$$

gdzie e reprezentuje zaobserwowane błędy, czyli reszty z dopasowania linii regresji $\widehat{\beta}_0 + \widehat{\beta}_1 X$ do zbioru obserwacji Y czyli

$$e = Y - \widehat{\beta}_0 - \widehat{\beta}_1 X$$

Dla poszczególnych par punktów mamy związek

$$\begin{aligned} y_i &= \widehat{\beta}_0 + \widehat{\beta}_1 x_i + e_i,\ i = 1, ..., n \\ e_i &= y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \end{aligned}$$

Linia regresji:

$$\begin{aligned} \widehat{Y} &= \widehat{\beta}_0 + \widehat{\beta}_1 X \\ \widehat{y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 x_i,\ i = 1, ..., n \end{aligned}$$

Wartości $\widehat{y}_i$ nazywa się estymatorami y_i, wartościami teoretycznymi lub prognozowanymi wartościami Y.

-Metoda najmniejszych kwadratów:

$$\begin{aligned} SSE &= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ SSE\left(\widehat{\beta}_0, \widehat{\beta}_1\right) &= \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 \\ & \min_{\widehat{\beta}_0, \widehat{\beta}_1} \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 \end{aligned}$$

Szukamy $\widehat{\beta}_0, \widehat{\beta}_1$ realizującymi minimum:

$$\frac{\partial SSE\left(\widehat{\beta}_0, \widehat{\beta}_1\right)}{\partial \widehat{\beta}_0} = 0$$

$$\frac{\partial SSE\left(\widehat{\beta}_0, \widehat{\beta}_1\right)}{\partial \widehat{\beta}_1} = 0$$

Ostatecznie

$$\begin{aligned} \widehat{\beta}_1 &= \frac{SS_{xy}}{SS_x} \\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x} \end{aligned}$$

gdzie

$$\begin{aligned} SS_x &= \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ SS_{xy} &= \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \end{aligned}$$

Uwaga 4 *Terminologia: układ równań normalnych:*

$$\begin{cases} 0 = \sum_{i=1}^{n} y_i - n\widehat{\beta}_0 - \widehat{\beta}_1 \sum_{i=1}^{n} x_i \\ 0 = \sum_{i=1}^{n} x_i y_i - \widehat{\beta}_0 \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 \end{cases}$$

Własności 4

1. $(\overline{x}, \overline{y})$ *należy do linii regresji*
2. *Układ równań normalnych jest równważny z*

$$\begin{cases} 0 = \sum_{i=1}^{n} e_i \\ 0 = \sum_{i=1}^{n} x_i e_i \end{cases}$$

3. $$\widehat{\beta}_1 = \sqrt{\frac{SS_y}{SS_x}} \widehat{\rho}_{XY}$$

4. $$\widehat{y}_i - \overline{y} = \widehat{\beta}_1 (x_i - \overline{x})$$

Uwaga 5 $\widehat{\beta}_0$ *i* $\widehat{\beta}_1$ *są wartościami (relizacjami) estymatorów prawdziwych parametrów regresji.*

1.3 Twierdzenie Gaussa-Markowa (przypadek jednowymiarowy)

Lemat 1

$$Var\left(\widehat{\beta}_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right), Var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{S_{XX}}$$

Twierdzenie 1 *Gaussa i Markowa*

W modelu regresji liniowej nieobciążonymi, liniowymi estymatorami parametrów β_0 *i* β_1 *o najmniejszej wariancji w klasie estymatorów liniowych są estymatory wyznaczone metodą najmniejszych kwadratów (**BLUE** =best linear unbiased estimator).*

Wniosek 1 $E\left(\widehat{\beta}_0\right) = \beta_0$, $E\left(\widehat{\beta}_1\right) = \beta_1$

Ćwiczenie 1

1. *Zaproponować (sztuczny) model regresji liniowej.*
2. *Wygenerować pary punktów z tego modelu.*
3. *Przeprowadzić estymację parametrów*
4. *Narysować wygenerowane pary punktów wraz z linią regresji*
5. *Interpretacja*

Solution 1 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, ..., 10$

1. *Narzędzia>Analiza Danych>Generowanie Liczb Pseudolosowych (normalny)*
2. *Zaznacz obszar 1x2, naciśnij F2 wpisz =REGLINP(Zakres_ Y;Zakres_ X;1;1); CTRL+SHIFT+ENTER*
3. *Na wykresie rozrzutu punktów (x_i, y_i) zaznaczyć punkty>Prawy Przycisk Myszy>Dodaj linię trendu.*
4. *Jeśli wartość zmiennej objaśniającej X zwiększy się o jednostkę, to wartość zmiennej Y zmieni się o $\widehat{\beta}_1$ (w zależności od znaku).*

Ćwiczenie 2 *Jak powyżej ale dla $n = 5000$. Wykorzystać generowanie liczb pseudolosowych dla X. Zbadaj wpływ wielkości odchylenia standardowego zmiennej ε na wyniki estymacji. Wnioski?*

Ćwiczenie 3 *Dla par $(x_i, y_i) = (x_i, x_{i+1})$, gdzie x_i jest i-tą obserwacją WIG20 przeprowadź estymację regresji (wyznacz estymatory, narysuj wykres rozrzutu wraz z linią regresji).*

1.4 Badanie istotności parametrów

Twierdzenie 2 *Jeśli ε mają rozkład normalny, to*

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{SSE}{n-2},$$

gdzie $SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$, jest nieobciążonym estymatorem parametru[1] *σ^2.*

Twierdzenie 3 *Jeśli ε mają rozkład normalny, to*

$$\frac{\widehat{\beta}_0 - \beta_0}{SE\left(\widehat{\beta}_0\right)} \sim t_{n-2}$$

$$\frac{\widehat{\beta}_1 - \beta_1}{SE\left(\widehat{\beta}_1\right)} \sim t_{n-2}$$

gdzie[2] $SE\left(\widehat{\beta}_0\right) = \sqrt{\widehat{\sigma}^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}$, $SE\left(\widehat{\beta}_1\right) = \sqrt{\frac{\widehat{\sigma}^2}{S_{XX}}}$.

[1] Dzielimy przez liczbę stopni swobody równą $n-2$. Odejmujemy 2 bo stopnie swobody zostały zwiazane przez estymatory współczynnków.

[2] $SE\left(\widehat{\beta}_0\right) = \sqrt{\widehat{Var\left(\widehat{\beta}_0\right)}}$, $SE\left(\widehat{\beta}_1\right) = \sqrt{\widehat{Var\left(\widehat{\beta}_1\right)}}$

1.4.1 Przedziały ufości

Lemat 2 *Jeśli ε mają rozkład normalny, to*

1. $(1-\alpha)\,100\%$ *przedział ufności dla parametru β_0 jest postaci*

$$\widehat{\beta}_0 \pm t_{n-2,\alpha/2} SE\left(\widehat{\beta}_0\right)$$

2. $(1-\alpha)\,100\%$ *przedział ufności dla parametru β_1 :*

$$\widehat{\beta}_1 \pm t_{n-2,\alpha/2} SE\left(\widehat{\beta}_1\right)$$

Uwaga 6 *Wskazówka praktyczna: jeśli wartość zero nie znajduje się w przedziale ufności parametru β_1 to z prawdopodobieństwem $1-\alpha$ zależność liniowa pomiędzy zmienną Y a zmienną X jest istotna (ściślej: współczynnik nachylenia linii regresji jest różny od zera).*

1.4.2 Testy istotności

Jeśli wartość zmiennej Y jest stała niezależnie od wartości X, to $\beta_1 = 0$.

Jeśli korelacja pomiędzy zmiennymi nie występuje to $\beta_1 = 0$.

Lemat 3 *Jeśli $\varepsilon \sim N(,)$ oraz $c \in \mathbb{R}$ to*

$$\begin{aligned} H_0 &: \beta_1 = c \\ H_0 &: \beta_1 \neq c \end{aligned}$$

$$\frac{\widehat{\beta}_1 - c}{SE\left(\widehat{\beta}_1\right)} \sim t_{n-2}$$

Jest to test dwustronny.

Wniosek 2 *Jeśli $\varepsilon \sim N(,)$ to*

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_0 &: \beta_1 \neq 0 \end{aligned}$$

$$\frac{\widehat{\beta}_1}{SE\left(\widehat{\beta}_1\right)} \sim t_{n-2}$$

Jest to test dwustronny.

Uwaga 7 *Gdy współczynnik korelacji jest istotnie różny od zera lub gdy istnieje związek regresyjny pomiędzy zmiennymi to i tak nie możemy stwierdzić związku przyczynowo-skutowego między zmiennymi tzn. nie możemy stwierdzić, że jedna zmienna jest przyczyną drugiej.*

Ćwiczenie 4 *Analiza regresji w oparciu o plik: PilkaNozna.xls.*

1. *Przeprowadzić estymację parametrów*
2. *Wyznaczyć błędy standardowe parametrów:* $SE\left(\widehat{\beta}_0\right) = \sqrt{\widehat{\sigma}^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)}$, $SE\left(\widehat{\beta}_1\right) = \sqrt{\frac{\widehat{\sigma}^2}{S_{XX}}}$
3. *Testuj:*

$$\begin{aligned} H0 &: \beta_1 = 0 \\ H1 &: \beta_1 \neq 0 \end{aligned}$$

Solution 2 *Zaznacz obszar 2x2, naciśnij F2 wpisz =REGLINP(Zakres_ Y;Zakres_ X**;1;1**) oraz CTRL+SHIFT+ENTER. Otrzymasz*

$$\begin{matrix} \widehat{\beta}_1 & \widehat{\beta}_0 \\ SE\left(\widehat{\beta}_1\right) & SE\left(\widehat{\beta}_0\right) \end{matrix}$$

$$t_{stat} = \frac{\widehat{\beta}_1}{SE\left(\widehat{\beta}_1\right)}$$

$$\begin{aligned} p-value &= ROZKŁAD.T(Moduł.Liczby(t_{stat}), n-2, 2) \\ U &= ROZKŁAD.T.ODW(\alpha/2; n-2) \\ Zbiór\ krytyczny &= (-\infty, -U) \cup (U, \infty) \end{aligned}$$

α *to prawdopodobieństwo zbioru krytycznego.*

$$\begin{aligned} ROZKŁAD.T(x, \nu, 2) &= P\left(|X| > x\right), gdzie X \sim t\left(\nu\right) \\ ROZKŁAD.T.ODW(p; \nu) &= U, gdzie\ P\left(X > U\right) = p \end{aligned}$$

Ćwiczenie 5 *Dla danych z WIG20 oraz ich logarytmów sprawdzić normalność. QQ-Plot: Chcemy ocenić czy* $x_1, ..., x_n$ *są realizacją rozkładu normalnego. Uporządkować próbkę* $x_1, ..., x_n$ *od najmniejszej do największej otrzymując* $x_{(1)}, ..., x_{(n)}$. *Zaznaczyć na układzie współrzędnym pary* $(x_{(j)}, z_j)$, *dla* $j = 1, ..., n$, *gdzie* z_j *wyznaczamy z tożsamości:*

$$\frac{j-0.5}{n} = N\left(\frac{z_j - \widehat{\mu}_n}{\widehat{\sigma}_n}\right)$$

gdzie N *jest dystrybuantą standardowego rozkładu normalnego* $N(0,1)$, $\widehat{\mu}_n$, $\widehat{\sigma}_n$ *są odpowiednio oszacowaniem wartości oczekiwanej* (μ) *i odchylenia standardowego* (σ) *z próby. Należy narysować „prostą regresji liniowej"dla* $(x_{(j)}, z_j)$. *Zaznaczone punkty nie powinny „zbytnioódbiegać od równania prostej.*

Ćwiczenie 6 *Zastosować regresję liniową do estymacji parametrów modelu dwumianowego*

$$\begin{aligned} S(k) &= S(k-1)\left(1+\xi_k U+(1-\xi_k)D\right) \\ S(0) &= S_0 \end{aligned}$$

$$\xi_k = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$$

$$\begin{aligned} Y_k &= \frac{S(k)}{S(k-1)} = 1+\xi_k U+(1-\xi_k)D \\ y_k &= Y_k - 1 = 1+\xi_k U+(1-\xi_k)D \\ &= D+\xi_k(U-D) \end{aligned}$$