Statystyczna analiza danych

Wykład 13

Analiza zależności dwóch zmiennych

Współczynnik korelacji próbkowej

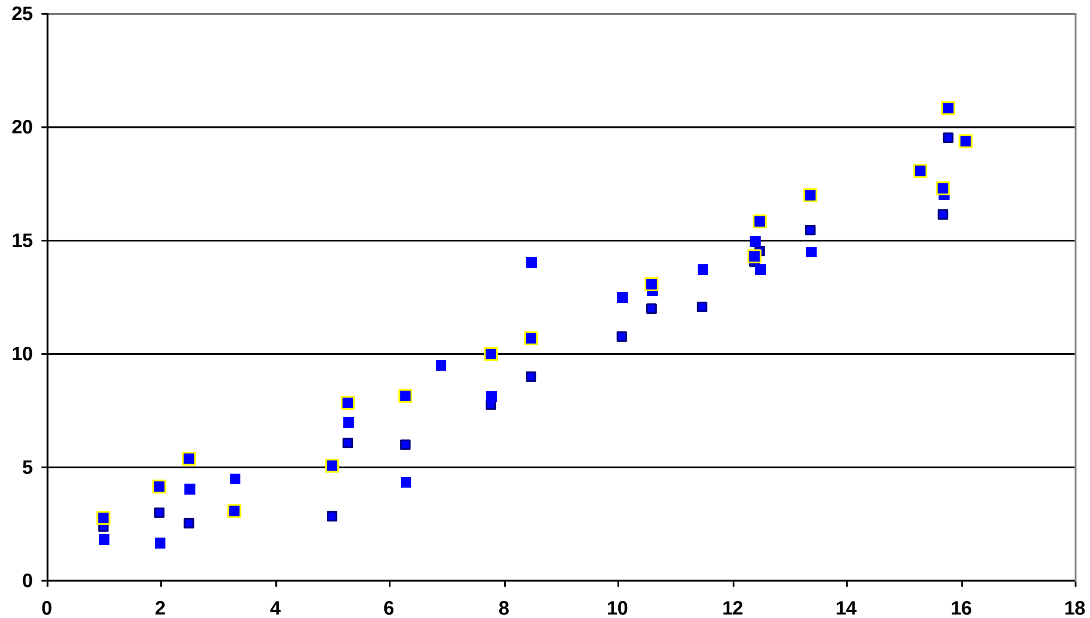
Niech $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ będzie próbką cechy dwuwymiarowej (X, Y).

Będziemy badali zależność Y od X.

X = zmienna **niezależna** (o**bjaśniająca**),

Y = zmienna **zależna** (**objaśniana**),

Wykres rozproszenia – graficzne przedstawienie próbki w postaci punktów na płaszczyźnie Oxy.



Współczynnik korelacji z próby

Definicja. Niech $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ będzie próbą losową.
Współczynnikiem korelacji z próby losowej nazywamy zmienną losową

$$\boxed{R=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{S_X}\right)\left(\frac{Y_i-\overline{Y}}{S_Y}\right)},$$

gdzie $\overline{X}$ i S_X oznaczają średnią i odchylenie standardowe dla $X_1,X_2,\ldots,X_n$,
a $\overline{Y}$ i S_Y oznaczają średnią i odchylenie standardowe dla $Y_1,Y_2,\ldots,Y_n$.

(np. $\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$, $S_Y^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2$, $S_Y=\sqrt{S_Y^2}$)

Współczynnik korelacji próbkowej

Współczynnikiem korelacji próbkowej nazywamy wartość współczynnika R obliczoną dla próbki $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$:

$$\boxed{r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_X}\right)\left(\frac{y_i - \overline{y}}{s_Y}\right)}$$

Własności współczynnika korelacji próbkowej

- $\boxed{-1 \le r \le 1}$.
- Jeśli $r = 1$, to wszystkie punkty wykresu rozproszenia <u>leżą na prostej o dodatnim współczynniku kierunkowym</u>, tzn. istnieje dodatnia zależność liniowa między zmiennymi *x* i *y* próbki.
- Jeśli $r = -1$, to wszystkie punkty wykresu rozproszenia <u>leżą na prostej o ujemnym współczynniku kierunkowym</u>, tzn. istnieje ujemna zależność liniowa między zmiennymi *x* i *y* próbki.
- Wartości *r* bliskie –1 lub 1 wskazują, że wykres rozproszenia <u>jest skupiony wokół pewnej prostej</u>.

Regresja prostoliniowa

Prosta regresji. Metoda najmniejszych kwadratów

Problem: W jaki sposób do wykresu rozproszenia, tzn. do punktów $(x_1, y_1),(x_2, y_2),\ldots,(x_n, y_n)$ dopasować „najlepiej" linię prostą?

Niech $\boxed{y = b_0 + b_1 x}$, $-\infty < x < \infty$, będzie równaniem prostej „dopasowanej" do punktów (x_i, y_i), $i = 1,\ldots,n$, wykresu rozproszenia.
(b_1 - **współczynnik kierunkowy,** b_0 - **wyraz wolny**).

Wówczas $\widehat{y}_i = b_0 + b_1 x_i$ będzie przybliżeniem wartości y_i na podstawie zmiennej niezależnej x_i uzyskanym z zależności liniowej.

Błąd przybliżenia, czyli różnicę $y_i - \widehat{y}_i$ nazywamy **wartością resztową,** lub **rezyduum.**

Regresja prostoliniowa

Miarą dopasowania prostej do próbki (punktów wykresu rozproszenia) jest suma kwadratów błędów (rezyduów):

$$\boxed{S(b_0,b_1)=\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2=\sum_{i=1}^{n}(y_i-(b_0+b_1x_i))^2}$$

Prostą dla której $S(b_0,b_1)$ osiąga wartość <u>minimalną</u> nazywamy **prostą regresji** lub też prostą wyznaczoną **metodą najmniejszych kwadratów.**

Współczynniki prostej regresji b_0, b_1 wyznaczamy z warunku koniecznego minimum funkcji $S(b_0,b_1)$, tzn. przyrównując do zera obie pochodne cząstkowe.

Regresja prostoliniowa

Rozwiązując układ dwóch równań liniowych otrzymujemy:

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

$$\boxed{b_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i\right) = \bar{y} - b_1 \bar{x} \qquad (2)}$$

gdzie $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Wartość $\boxed{\hat{y} = b_0 + b_1 x}$ nazywamy **wartością przewidywaną** zmiennej objaśnianej (zależnej), przy pomocy prostej regresji, na podstawie zmiennej objaśniającej (niezależnej) *x*.

Regresja prostoliniowa

Ocena „dobroci" dopasowania prostej regresji

Wprowadzamy oznaczenia:

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

- **całkowita suma kwadratów** (*Total Sum of Squares*) (miara zmienności samych $y_1, \ldots, y_n$).

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

- **suma kwadratów błędów** (*Error Sum of Squares*).

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

- **regresyjna (modelowa) suma kwadratów** (*Regression/Model Sum of Squares*) (miara zmienności $\hat{y}_1, \ldots, \hat{y}_n$).

Regresja prostoliniowa

Można pokazać, że zachodzi równość :

$$\boxed{\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}.$$

$$SST \quad = \quad SSE \quad + \quad SSR$$

Współczynnik determinacji określony wzorem

$$\boxed{R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}}$$

jest miarą stopnia dopasowania prostej regresji do wykresu rozproszenia.

Określa stopień, w jakim zależność liniowa, między zmienną objaśnianą a objaśniającą, wyjaśnia zmienność wykresu rozproszenia.

Im mniejsze SSE, tym wykres rozproszenia jest bardziej skupiony wokół prostej regresji.

Regresja prostoliniowa

Wartość współczynnika determinacji jest ściśle związana z wartością współczynnika korelacji próbkowej.

Stwierdzenie.

Zachodzi równość

$$\boxed{r^2 = \frac{SSR}{SST} = R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

zmienność wyjaśniona przez model/ zmienność całkowita

Regresja prostoliniowa

Przykład. Zanotowano miesięczne wydatki na reklamę (w 10 000 PLN) pewnego artykułu oraz miesięczne dochody ze sprzedaży artykułu (w 100 000 PLN)

Miesiąc i	1	2	3	4	5
Reklama x_i	5	6	7	8	9
Dochód y_i	4,5	6,5	8,4	7,6	8,4

Wyznaczyć liniową funkcję regresji oraz przewidywaną wartość dochodu przy wydatkach na reklamę 100 000 PLN (10 x 10 000).

Kolejno obliczamy:

$$\overline{x} = 7{,}0, \quad \overline{y} = 7{,}08, \quad s_X = 1{,}58, \quad s_Y = 1{,}64$$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_X}\right)\left(\frac{y_i - \overline{y}}{s_Y}\right) = 0{,}858$$

(współczynnik korelacji próbkowej)

Regresja prostoliniowa

Współczynniki prostej regresji

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \underline{0{,}89}$$

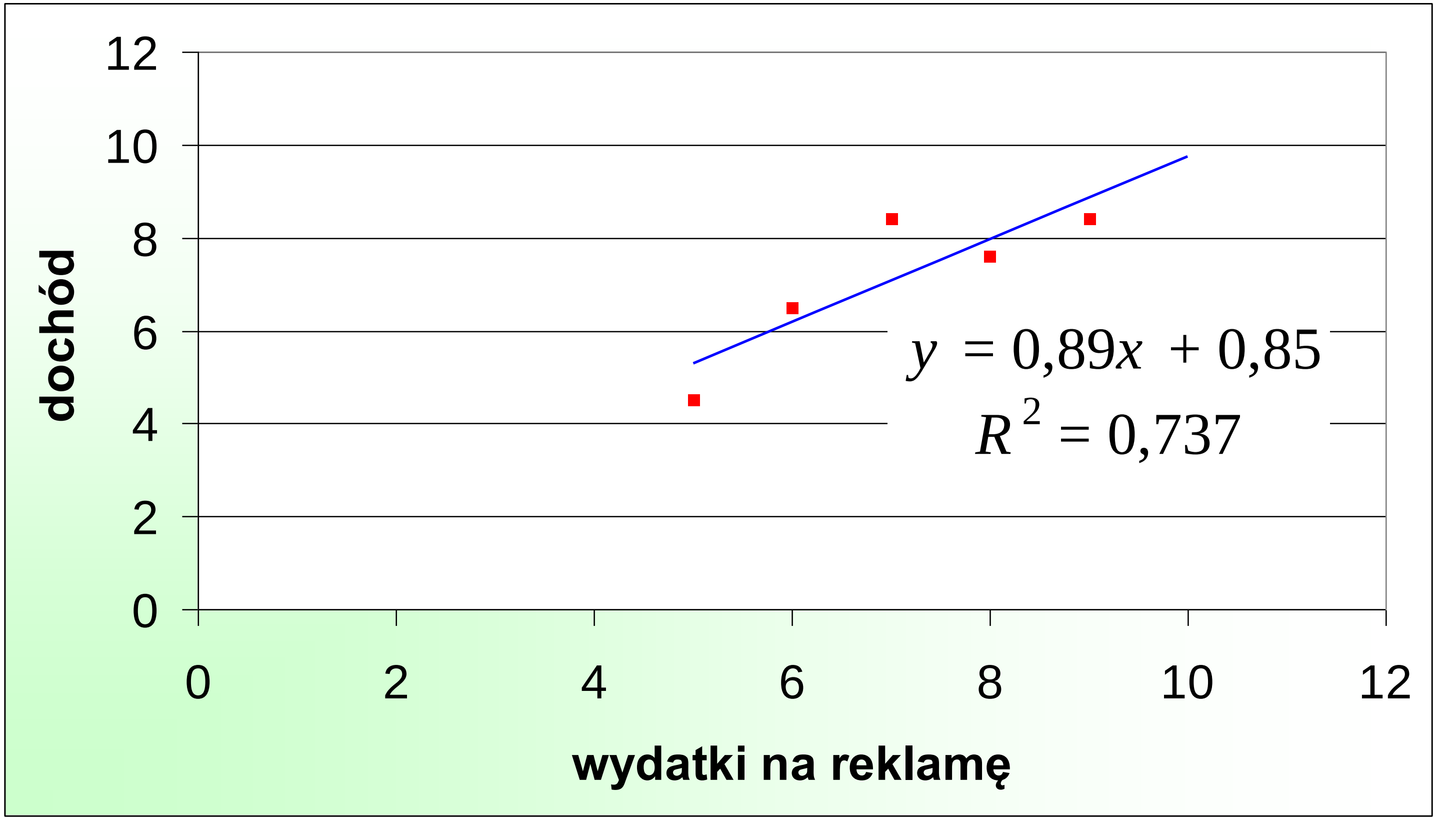
$$b_0 = \bar{y} - b_1\bar{x} = 7{,}08 - 0{,}89 \times 7 = \underline{0{,}85}$$

Przewidywany dochód ze sprzedaży, przy wydatku na reklamę x = 10 (x 10 000 PLN)

$$\hat{y} = b_0 + b_1 x = 0{,}85 + 0{,}89 \times 10 = 9{,}75\,(\times 10\ 000\ \text{PLN})$$

Regresja prostoliniowa

Wykres rozproszenia oraz empiryczna prosta regresji



Regresja prostoliniowa

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 10{,}748$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 2{,}827$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 7{,}921$$

współczynnik determinacji $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

$$R^2 = 0{,}737$$

Zmienność dochodu jest w prawie 74% wyjaśniona przez zmienność wydatków ma reklamę (zmienność wydatków na reklamę w 74% określa zmienność dochodu).

Regresja prostoliniowa

Model zależności liniowej (**model regresji liniowej**)

Załóżmy, że próbka $(x_1, y_1),\ldots,(x_n, y_n)$ jest realizacją próby losowej $(x_1, Y_1),\ldots,(x_n, Y_n)$, gdzie

$$\boxed{Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}, \quad i = 1,\ldots,n,$$

oraz $\varepsilon_1, \varepsilon_2,\ldots,\varepsilon_n$ są <u>niezależnymi</u> zmiennymi losowymi o wartości oczekiwanej 0 i wariancji σ^2, a znane liczby $x_1,\ldots,x_n$ nie wszystkie są jednakowe.

Prostą $\boxed{y = \beta_0 + \beta_1 x}$ nazywamy **prostą regresji.**

Współczynnik β_0 - **wyraz wolny** prostej regresji.

Współczynnik β_1 - **współczynnik kierunkowy** prostej regresji.

Zmienne losowe $\varepsilon_1, \varepsilon_2,\ldots,\varepsilon_n$ - **losowe błędy** w modelu $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

Regresja prostoliniowa

Własności zmiennej losowej Y_i, $i = 1,\dots,n$,

$$E(Y_i) = E(\beta_0 + \beta_1 x_i) + E(\varepsilon_i) = \beta_0 + \beta_1 x_i \,.$$

$$\mathrm{Var}(Y_i) = \mathrm{Var}(\beta_0 + \beta_1 x_i + \varepsilon_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$$

Założenia:

- $x_1,\dots,x_n$ są znane.
- Obserwujemy wartości zmiennych $Y_1,\dots,Y_n$.
- $\beta_0, \beta_1, \sigma^2$ są nieznanymi parametrami modelu.

Cel eksperymentu – wnioskowanie na temat parametrów modelu

Regresja prostoliniowa

Naturalne **estymatory** parametrów β_0, β_1 otrzymujemy metodą najmniejszych kwadratów, wstawiając we wzorach (1), (2) zmienne losowe Y_i zamiast ich wartości y_i, $i = 1,\ldots,n$,

$$\boxed{\widehat{b_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}.$$

$$\boxed{\widehat{b_0} = \overline{Y} - \widehat{b_1}\overline{x}}.$$

Regresja prostoliniowa

Własności estymatorów $\widehat{b}_0, \widehat{b}_1$ podane są w następującym twierdzeniu.

Twierdzenie

(i) $E(\widehat{b}_0) = \beta_0, \qquad E(\widehat{b}_1) = \beta_1$

(ii)
$$\mathrm{Var}(\widehat{b}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right) \qquad (3)$$

(iii)
$$\mathrm{Var}(\widehat{b}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \qquad (4)$$

(iv) Jeśli $\varepsilon_i \sim N(0,\sigma)$, $i = 1,..,n$, to $\widehat{b}_0, \widehat{b}_1$ mają rozkłady normalne o wartościach średnich i wariancjach określonych w **(i) - (iii).**

Regresja prostoliniowa

Estymator σ^2

Definicja

Błędem średniokwadratowym S^2 nazywamy estymator wariancji σ^2 określony następująco

$$\boxed{S^2 = \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n-2} = \frac{SSE}{n-2}}$$

Liczbę $n-2$ nazywamy liczbą stopni swobody rezyduów.

Stwierdzenie

S^2 **jest nieobciążonym estymatorem** σ^2**, tzn.**

$$E(S^2) = \sigma^2.$$

$S = \sqrt{S^2}$ = estymator σ.

Regresja prostoliniowa

Wniosek (i) Nieobciążonym estymatorem wariancji $\mathrm{Var}(\widehat{b}_0)$ jest

$$[SE(\hat{b}_0)]^2 = S^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right),$$

Stąd naturalnyym estymatorem odchylenia standardowego $\sigma_{\widehat{b}_0}$ jest

$$SE(\widehat{b}_0) = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

nazywany **błędem standardowym estymatora** $\widehat{b}_0$, gdyż na mocy (3)

$$SE(\widehat{b}_0) = \text{ estymator } \sigma_{\widehat{b}_0} = \sqrt{Var(\widehat{b}_0)}$$

Regresja prostoliniowa

(ii) Nieobciążonym estymatorem $\mathrm{Var}(\widehat{b}_1)$ jest

$$\boxed{[SE(\widehat{b}_1)]^2 = \frac{S^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}},$$

Stąd naturalnym estymatorem odchylenia standardowego $\sigma_{\widehat{b}_1}$ jest

$$\boxed{SE(\widehat{b}_1) = \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}}$$

nazywamy **błędem standardowym** estymatora $\widehat{b}_1$, gdyż na mocy (4)

$$SE(\widehat{b}_1) = \text{ estymator } \sigma_{\widehat{b}_1} = \sqrt{Var(\widehat{b}_1)}.$$

Regresja prostoliniowa

Twierdzenie Jeśli $\varepsilon_i \sim N(0,\sigma)$, $i = 1,..,n$, to:

(i)

$$\boxed{\widehat{b}_1 \sim N\left(\beta_1, \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right)},$$

(tzn. ma rozkład normalny ze wskazanymi parametrami),

$$\boxed{\frac{\widehat{b}_1 - \beta_1}{SE(\widehat{b}_1)} \sim t_{n-2}}$$

gdzie

$$SE(\widehat{b}_1) = \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\ .$$

(tzn. ma rozkład Studenta o n-2 stopniach swobody).

Regresja prostoliniowa

(ii) $$\boxed{\widehat{b}_0 \sim N\left(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\right)},$$

(tzn. ma rozkład normalny ze wskazanymi parametrami),

$$\boxed{\frac{\widehat{b}_0 - \beta_0}{SE(\widehat{b}_0)} \sim t_{n-2}}$$

gdzie $$SE(\widehat{b}_0) = S\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

(tzn. ma rozkład Studenta o n-2 stopniach swobody).

Regresja prostoliniowa

Przedział ufności na poziomie ufności $1-\alpha$ dla **współczynnika** β_1:

$$\boxed{[\widehat{b}_1 - t_{1-\alpha/2,n-2} \times SE(\widehat{b}_1), \quad \widehat{b}_1 + t_{1-\alpha/2,n-2} \times SE(\widehat{b}_1)]}$$

Przedział ufności na poziomie ufności $1-\alpha$ dla **współczynnika** β_0:

$$\boxed{[\widehat{b}_0 - t_{1-\alpha/2,n-2} \times SE(\widehat{b}_0), \quad \widehat{b}_0 + t_{1-\alpha/2,n-2} \times SE(\widehat{b}_0)]}$$

Regresja prostoliniowa

Testowanie hipotezy o wartości współczynnika β_0

(A) $\boxed{H_0 : \beta_0 = \beta_{0,0}}$,

gdzie $\beta_{0,0}$ jest ustaloną liczbą.

Statystyka testowa

$$\boxed{T = \frac{\widehat{b}_0 - \beta_{0,0}}{SE(\widehat{b}_0)} = (\widehat{b}_0 - \beta_{0,0})/(S\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}})}$$

Jeśli H_0 **prawdziwa**, to $T \sim t_{n-2}$.

(ma rozkład Studenta o n-2 stopniach swobody).

Regresja prostoliniowa

Zbiory krytyczne dla różnych postaci hipotez alternatywnych

(a) $\boxed{H_1 : \beta_0 \neq \beta_{0,0}}$.

Zbiór krytyczny $C = \{t : |t| \geq t_{1-\alpha/2,n-2}\}$.

(b) $\boxed{H_1 : \beta_0 > \beta_{0,0}}$.

Zbiór krytyczny $C = \{t : t \geq t_{1-\alpha,n-2}\}$.

(c) $\boxed{H_1 : \beta_0 < \beta_{0,0}}$

Zbiór krytyczny $C = \{t : t \leq -t_{1-\alpha,n-2}\}$.

Regresja prostoliniowa

Testowanie hipotezy o wartości współczynnika β_1

(B) $\boxed{H_0 : \beta_1 = \beta_{1,0}}$,

gdzie $\beta_{1,0}$ jest ustaloną liczbą.

Statystyka testowa

$$\boxed{T = \frac{\widehat{b_1} - \beta_{1,0}}{SE(\widehat{b_1})} = \frac{(\widehat{b_1} - \beta_{1,0})\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}{S}}$$

Jeśli H_0 **prawdziwa**, to $T \sim t_{n-2}$.

(ma rozkład Studenta o n-2 stopniach swobody).

Regresja prostoliniowa

Zbiory krytyczne dla różnych postaci hipotez alternatywnych

(a) $\boxed{H_1 : \beta_1 \neq \beta_{1,0}}$.

Zbiór krytyczny $C = \{t : |t| \geq t_{1-\alpha/2,n-2}\}$.

(b) $\boxed{H_1 : \beta_1 > \beta_{1,0}}$.

Zbiór krytyczny $C = \{t : t \geq t_{1-\alpha,n-2}\}$.

(c) $\boxed{H_1 : \beta_1 < \beta_{1,0}}$.

Zbiór krytyczny $C = \{t : t \leq -t_{1-\alpha,n-2}\}$.

Regresja prostoliniowa

c) $\boxed{H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0}$,

Statystyka testowa

$$\boxed{F = \frac{SSR/1}{SSE/(n-2)}}$$

Jeśli H_0 prawdziwa, to F ma **rozkład** F **Snedecora** o $(1,n\text{-}2)$ stopniach swobody.

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 .$$

SST	=	SSE	+	SSR
n-1	=	n-2	+	1

(Liczby stopni swobody SSx = liczba niezależnych zmiennych zmniejszona o liczbę ograniczeń występujących w określeniu SSx).

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i , \quad i = 1,\ldots,n$$

Regresja prostoliniowa

Zbiór krytyczny testu

$$\boxed{C = \{F_{obl} : F_{obl} \geq f_{1-\alpha,1,n-2}\}}.$$

Zauważmy, że

$$\boxed{F = T^2},$$

stąd test jest szczególnym przypadkiem testu z **(B)** gdy $\beta_{1,0} = 0$.