

Elementarna statystyka

Przedział przewidywanej wartości dla zmiennej Y (Prediction interval for unobserved Y)

Alexander Bendikov

Uniwersytet Wrocławski

2 czerwca 2016

Przedział przewidywanej wartości dla zmiennej

Przykład: Dział rekrutacji pewnej uczelni chciałby przewidzieć Y – średnią ocen studenta po pierwszym roku studiów na podstawie X – średniej ocen tego studenta w szkole średniej. Uczelnia ma dane dotyczące dawnych studentów, zawierające ich średnie ocen w szkole średniej i po pierwszym roku studiów. Takie dane mogą być użyte do prognozy średniej ocen po pierwszym roku nowych studentów.

Przedział przewidywanej wartości dla zmiennej

Przykład: Dział rekrutacji pewnej uczelni chciałby przewidzieć Y – średnią ocen studenta po pierwszym roku studiów na podstawie X – średniej ocen tego studenta w szkole średniej. Uczelnia ma dane dotyczące dawnych studentów, zawierające ich średnie ocen w szkole średniej i po pierwszym roku studiów. Takie dane mogą być użyte do prognozy średniej ocen po pierwszym roku nowych studentów.

Założenia: Mamy dwie próby: $X : x_1, x_2, \dots$ (średnie ze szkoły średniej) oraz odpowiadającą jej $Y : y_1, y_2, \dots$ (średnie po pierwszym roku studiów). Mamy interesującą nas zmienną Y , odpowiadającą $X = x^*$. Zakładamy:

- Y jest niezależna od $y_1, y_2, \dots$.
- $Y \in N(\alpha + \beta x^*, \sigma)$.

Mamy więc:

- $Y - \hat{Y} = Y - (a + bx^*),$
- $E(Y - \hat{Y}) = \alpha + \beta x^* - (E a + x^* E b) = 0,$
- $Var(Y - \hat{Y}) = Var(Y) + Var(\hat{Y}) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right).$

Mamy więc:

- $Y - \hat{Y} = Y - (a + bx^*),$
- $E(Y - \hat{Y}) = \alpha + \beta x^* - (E a + x^* E b) = 0,$
- $Var(Y - \hat{Y}) = Var(Y) + Var(\hat{Y}) = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right).$

Uwaga: Skoro σ^2 jest nieznana, zastępujemy ją przez s^2 i używamy następującej statystyki:

$$T = \frac{Y - \hat{Y}}{s \sqrt{1 + \frac{1}{n} \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}} = \frac{Y - \hat{Y}}{SE_Y}.$$

Mamy więc:

- $Y - \hat{Y} = Y - (a + bx^*),$
- $E(Y - \hat{Y}) = \alpha + \beta x^* - (E a + x^* E b) = 0,$
- $Var(Y - \hat{Y}) = Var(Y) + Var(\hat{Y}) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right).$

Uwaga: Skoro σ^2 jest nieznana, zastępujemy ją przez s^2 i używamy następującej statystyki:

$$T = \frac{Y - \hat{Y}}{s \sqrt{1 + \frac{1}{n} \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}} = \frac{Y - \hat{Y}}{SE_Y}.$$

Skoro s i $Y - \hat{Y}$ są niezależne, T ma rozkład Studenta $t(n - 2)$. Możemy więc utworzyć przedział ufności na poziomie C :

$$\hat{Y} \pm t^* SE_Y,$$

gdzie

$$SE_Y^2 = s^2 + SE_{\mu}^2 = s^2 \left(1 + \frac{1}{n} \right) + (x^* - \bar{x})^2 SE_b^2.$$

Uwaga: Wartości s i SE_b można wyliczyć z danych, więc z powyższego wzoru otrzymujemy SE_μ oraz SE_Y .

Uwaga: Wartości s i SE_b można wyliczyć z danych, więc z powyższego wzoru otrzymujemy SE_μ oraz SE_Y .

Przykład: Wróćmy do przykładu autokomisu z 62 Mazdami. Przypomnijmy wyniki wygenerowane na podstawie danych (wiek i cena auta) z programu Minitab:

$$\overline{\text{wiek}} = 7,63$$

$$\text{odch. stand. wieku} = 4,30$$

$$\overline{\text{cena}} = 10283,1$$

$$\text{och. stand. ceny} = 7922,3$$

$$r^2 = 0,677$$

$$s = 4540,97$$

Otrzymane współczynniki:

Zmienna	Wartość	Odch. stand.
a	21833,6	1180,8
b	-1514,0	135,1

Pytanie: Właściciel zaproponował klientce samochód, którego oczekuje lada moment (nie ma go wśród danych analizowanych powyżej). Wiadomo, że oczekiwany samochód ma 7 lat, ale właściciel komisju nie widział go jeszcze, i nie wycenił). Chciałby jednak dać klientce pewne oszacowanie ceny, na podstawie powyższych danych. Podsumowując: wyznaczmy 90% przedział ufności dla przewidywanej ceny oczekiwanej Mazdy.

Pytanie: Właściciel zaproponował klientce samochód, którego oczekuje lada moment (nie ma go wśród danych analizowanych powyżej). Wiadomo, że oczekiwany samochód ma 7 lat, ale właściciel komisju nie widział go jeszcze, i nie wycenił). Chciałby jednak dać klientce pewne oszacowanie ceny, na podstawie powyższych danych. Podsumowując: wyznaczmy 90% przedział ufności dla przewidywanej ceny oczekiwanej Mazdy.

Nasze oszacowanie ceny Mazdy na podstawie linii regresji:

$\hat{y} = 21833,6 + (-1514) \cdot 7 = 11235,6$. Nas interesuje cena konkretnego auta, a nie średnia, więc wyznaczamy 90% przedział przewidywanej wartości, $\hat{y} \pm t^* SE_y$, gdzie $t^* = t^*(62 - 2) = 1,671$ oraz

$$SE_y = \sqrt{s^2 + SE_\mu^2}.$$

Pytanie: Właściciel zaproponował klientce samochód, którego oczekuje lada moment (nie ma go wśród danych analizowanych powyżej). Wiadomo, że oczekiwany samochód ma 7 lat, ale właściciel komisju nie widział go jeszcze, i nie wycenił). Chciałby jednak dać klientce pewne oszacowanie ceny, na podstawie powyższych danych. Podsumowując: wyznaczmy 90% przedział ufności dla przewidywanej ceny oczekiwanej Mazdy.

Nasze oszacowanie ceny Mazdy na podstawie linii regresji:

$\hat{y} = 21833,6 + (-1514) \cdot 7 = 11235,6$. Nas interesuje cena konkretnego auta, a nie średnia, więc wyznaczamy 90% przedział przewidywanej wartości, $\hat{y} \pm t^* SE_y$, gdzie $t^* = t^*(62 - 2) = 1,671$ oraz

$SE_y = \sqrt{s^2 + SE_\mu^2}$. Mamy:

$$SE_\mu = \sqrt{\frac{s^2}{n} + (x^* - \bar{x})^2 SE_b^2} = \sqrt{\frac{4540,97^2}{62} + (7 - 7,63)^2 135,1^2} \\ \approx 582,95 \Rightarrow \text{C.I.}(\mu): 11235,6 \pm 974,1$$

$$SE_y = \sqrt{4540,97^2 + 582,95^2} \approx 4578,2 \\ \Rightarrow \text{C.I.}(y): 1123,6 \pm 7650,2,$$

czyli

$$3585 < y < 18886.$$

czyli

$$3585 < y < 18886.$$

Przykład: Poniżej mamy dane dotyczące ilości lat kształcenia (zmienna X) oraz rocznego dochodu (zmienna Y) dla próby $n = 18$ wybranej spośród 40-letnich mężczyzn.

X	10	16	12	6	12	12	16	16	18	12	10	12	16	14	11	12	19	16
Y	48	58	36	33	45	50	55	47	48	48	46	41	54	50	41	47	49	44

czyli

$$3585 < y < 18886.$$

Przykład: Poniżej mamy dane dotyczące ilości lat kształcenia (zmienna X) oraz rocznego dochodu (zmienna Y) dla próby $n = 18$ wybranej spośród 40-letnich mężczyzn.

X	10	16	12	6	12	12	16	16	18	12	10	12	16	14	11	12	19	16
Y	48	58	36	33	45	50	55	47	48	48	46	41	54	50	41	47	49	44

Na podstawie tych danych programy obliczył:

Współczynnik	SE	t -wartość
30,84249	5,103363	2,124577
1,18681	0,372311	3,187693
$r^2 = 0,3884116$	$df = 18 - 2 = 16$	
$s = 5,02275$		

Typowe pytania:

- a) Jaki % obserwowanej zmienności w dochodzie może być wytłumaczony linią regresji pomiędzy wykształceniem i dochodem?
- b) Czy dane dostarczają mocnego dowodu na liniową zależność pomiędzy wykształceniem i dochodem?
- c) Przyjrzyjmy się konkretnemu 40-latkowi, który uczęszczał do szkół przez 16 lat.
 - (i) Podaj 95% przedział ufności dla jego średniego dochodu
 - (ii) Podaj 95% przedział ufności dla jego konkretnego dochodu

Typowe pytania:

- a) Jaki % obserwowanej zmienności w dochodzie może być wytłumaczony linią regresji pomiędzy wykształceniem i dochodem?
- b) Czy dane dostarczają mocnego dowodu na liniową zależność pomiędzy wykształceniem i dochodem?
- c) Przyjrzyjmy się konkretnemu 40-latkowi, który uczęszczał do szkół przez 16 lat.
 - (i) Podaj 95% przedział ufności dla jego średniego dochodu
 - (ii) Podaj 95% przedział ufności dla jego konkretnego dochodu

Odpowiedzi:

- a) $r^2 = 0,388$, a więc zależność liniowa wyjaśnia 38,8% obserwowanej zmienności dochodów.
- b)
 - Formułujemy hipotezy: $H_0 : \beta = 0$ (brak zależności), $H_a : \beta \neq 0$.
 - $T = \frac{b}{SE_b} = 3,187693$, $df = 16$, $0,0025 < p < 0,005$, więc $0,005 < P < 0,01$
 - Mamy więc mocny dowód na to, że zależność istnieje.

c) (i)

$$\hat{y} = 30,84249 + 1,186813x^*|_{x^*=16} = 49,831$$

$$\begin{aligned} SE_{\mu} &= \sqrt{\frac{s^2}{n} + (x^* - \bar{x})^2 SE_b^2} = (\bar{x} = 13,33) = \\ &= \sqrt{\frac{5,02275^2}{18} + (16 - 13,33)^2 \cdot 0,372311^2} = 1,5451. \end{aligned}$$

Mamy $t^* = 2,120$ ($df = 16$, $C = 95\%$), więc przedział ma postać:

$$\hat{y} \pm t^* SE_{\mu} = 49,831 \pm (2,120)(1,5451) = 49,831 \pm 3,276,$$

czyli przedział ma postać

$$46,555 < \mu_y < 53,107.$$

c) (i)

$$\hat{y} = 30,84249 + 1,186813x^*|_{x^*=16} = 49,831$$

$$\begin{aligned} SE_{\mu} &= \sqrt{\frac{s^2}{n} + (x^* - \bar{x})^2 SE_b^2} = (\bar{x} = 13,33) = \\ &= \sqrt{\frac{5,02275^2}{18} + (16 - 13,33)^2 \cdot 0,372311^2} = 1,5451. \end{aligned}$$

Mamy $t^* = 2,120$ ($df = 16$, $C = 95\%$), więc przedział ma postać:

$$\hat{y} \pm t^* SE_{\mu} = 49,831 \pm (2,120)(1,5451) = 49,831 \pm 3,276,$$

czyli przedział ma postać

$$46,555 < \mu_y < 53,107.$$

(ii)

$$SE_y = \sqrt{s^2 + SE_{\mu}^2} = \sqrt{5,02275^2 + 1,5451^2} = 5,2550$$

$$\begin{aligned} \hat{y} \pm t^* SE_y &= 49,831 \pm 2,120 \cdot 5,2550 = \\ &= 49,832 \pm 11,1406 \end{aligned}$$

Przedział przewidywanej wartości: