Komputerowa analiza szeregów czasowych - raport 1

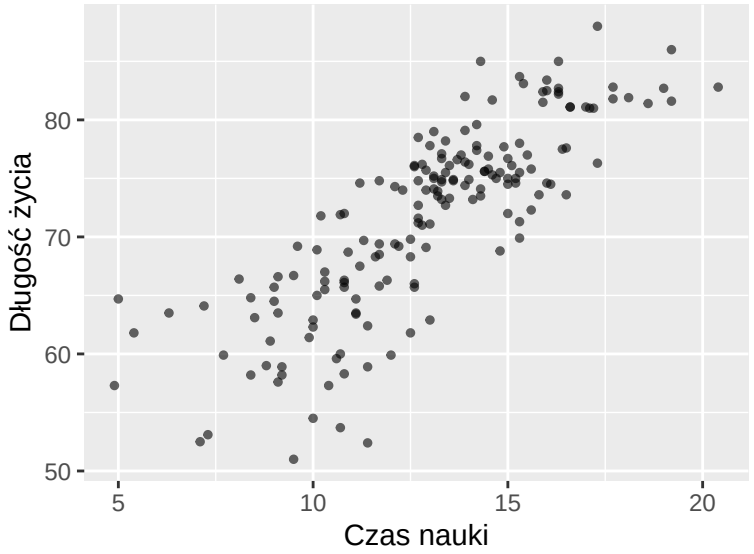
Szymon Malec, Tomasz Hałas

Wstęp

Celem raportu jest zbadanie liniowej korelacji pomiędzy edukacją, a długością życia. Wykorzystane do tego zostaną dane dostępne pod linkiem. Przeprowadzimy dokładną analizę zależności między dwoma zmiennymi oraz zastosujemy odpowiednie metody, by dopasować prostą regresji do danych, po czym zweryfikujemy czy dopasowanie można uznać za prawidłowe. Otrzymane wyniki wykorzystane zostaną do przeprowadzenia predykcji długości życia w zależności od czasu edukacji.

Opis danych

Dane, z których będziemy korzystać zawierają dużo informacji o poszczególnych 183 krajach, pozyskanych w latach 2000–2015, pozwalających okreslić ich aktualny rozwój lub możliwą długość życia mieszkanców. W naszym raporcie skorzystamy wyłącznie z oczekiwanej długości życia, średniego czasu edukacji oraz populacji w 2015 roku, czyli najbardziej aktualnych danych. Dla tego roku pełna informacja o pierwszych dwóch zmiennych zawarta jest dla 173 krajów.



Wykres 1: Wykres punktowy zależności pomiędzy czasem edukacji, a długością życia.

Na powyższym wykresie widzimy wyraźną zależność liniową pomiędzy dwoma zmiennymi. W dalszej części poddamy ją głębszej analizie.

Statystyki

Dla badanych zmiennych, czyli oczekiwanej długości życia i średniego czasu edukacji, obliczyliśmy najważniejsze statystyki. Wyniki przedstawiliśmy w poniższej tabeli:

	Długość życia	Czas edukacji
Średnia	71.71	12.93
Mediana	73.90	13.10
Odch. stand.	2.91	7.93
Minimum	51.00	4.90
Maksimum	88.00	20.40
IQR	10.70	4.20

Tablica 1: Najważniejsze statystyki obliczone dla badanych danych.

Widzimy, że oczekiwana długość życia wynosi średnio około 72 lata. Natomiast czas nauczania odznacza się dużą rozbieżnością. Chcąc jednak otrzymać średnią długość życia na świecie, musimy uwzględnić to że, dla każdego kraju jego oczekiwana długość życia przypada na każdego obywatela, a badane państwa mają różne populacje. Nie możemy zatem traktować ich jednakowo. Z tego powodu skorzystamy ze średniej ważonej, gdzie wagami będą populacje. Otrzymaliśmy wynik równy 70.13, przy czym należy zaznaczyć, że jest to wynik dla 142 państw, ponieważ kolumna zawierająca dane o populacji zawiera więcej brakujących danych.

Regresja

Aby dopasować do danych prostą regresji, skorzystamy z metody najmniejszych kwadratów. Oznaczmy licznę danych (państw) jako $n = 173$. Przyjmijmy model

$$Y_i = \beta_1 x_i + \beta_0 + \epsilon_i, \quad i = 1, 2, \ldots, n$$

gdzie x_i to dane dotyczące czasu nauczania, a ϵ_i są niezależnymi zmiennymi losowymi ze średnią równą 0 i skończoną wariancją. Oznaczmy dane z czasem życia jako y_i - będziemy traktować je jako realizacje zmiennych losowych Y_i. Wspomniana metoda polega na znalezieniu takich współczynników β_1, β_0, dla których funkcja

$$S(\beta_1, \beta_0) = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$$

przyjmuje wartość najmniejszą. Rozwiązaniem jest para estymatorów

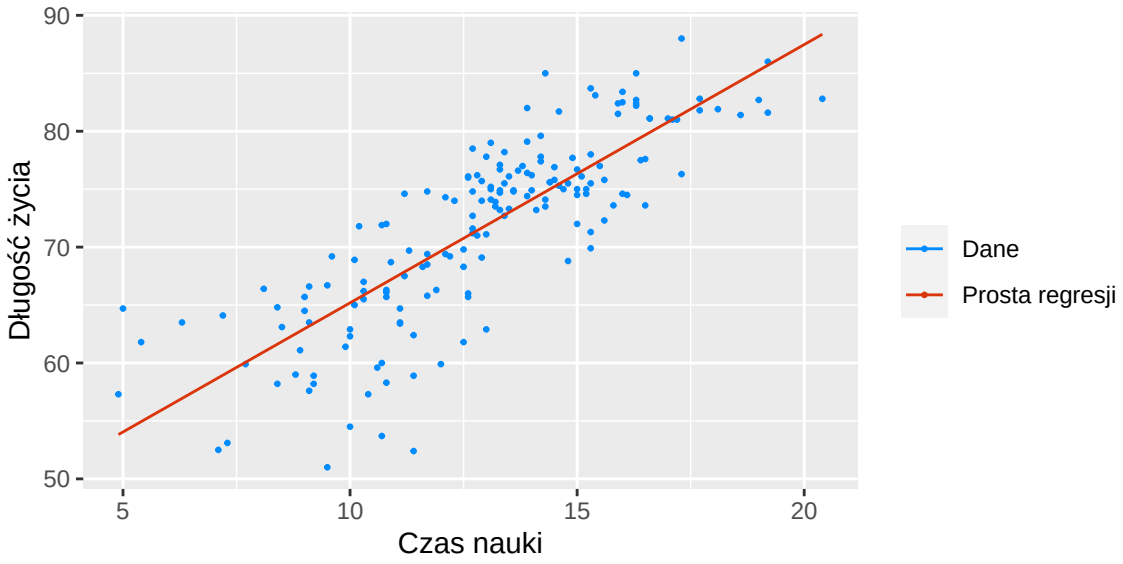
$$\begin{cases} \hat{\beta}_1 = R\dfrac{S_y}{S_x} = \dfrac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

gdzie R jest współczynnikiem korelacji Pearsona, a S_x, S_y są próbkowymi odchyleniami standardowymi. Można pokazać, że estymatory te są nieobciążone. Po podstawieniu danych otrzymujemy

$$\begin{cases} \hat{\beta}_1 \approx 2.23 \\ \hat{\beta}_0 \approx 42.9 \end{cases}.$$

Wyestymowane wartości Y_i będą miały postać

$$\hat{y_i} = \hat{\beta}_1 x_i + \hat{\beta}_0.$$



Wykres 2: Wykres punktowy wraz z prostą regresji wyznaczoną dla danych.

Jak możemy zauważyć, wyznaczona prosta pokrywa się z danymi. Współczynnik determinancji, czyli R^2 wyniósł 0.67, a więc jest to dość zadowalające dopasowanie.

Analiza residuów

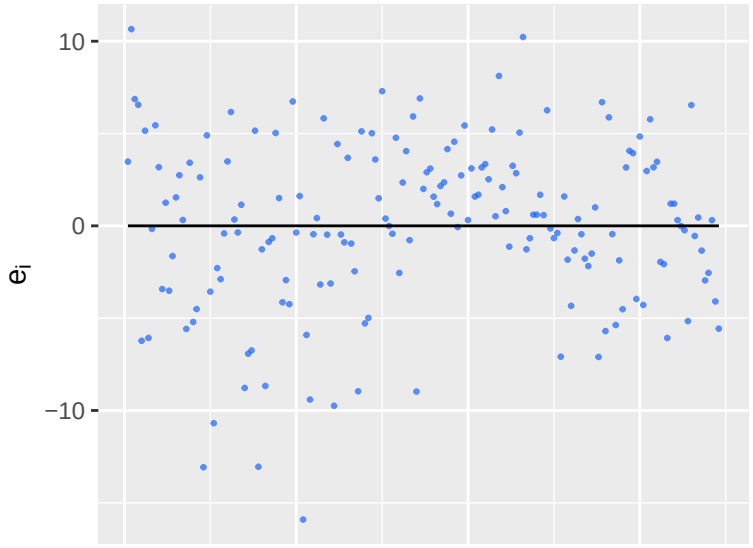
Aby sprawdzić, czy dane spełniają założenia modelu tj.

1. $\mathrm{E}\epsilon_i = 0$,
2. $\mathrm{Var}(\epsilon_i) < \infty$,
3. ϵ_i są niezależne,

przeprowadzimy analizę residuów (błędów)

$$e_i = y_i - \hat{y}_i,$$

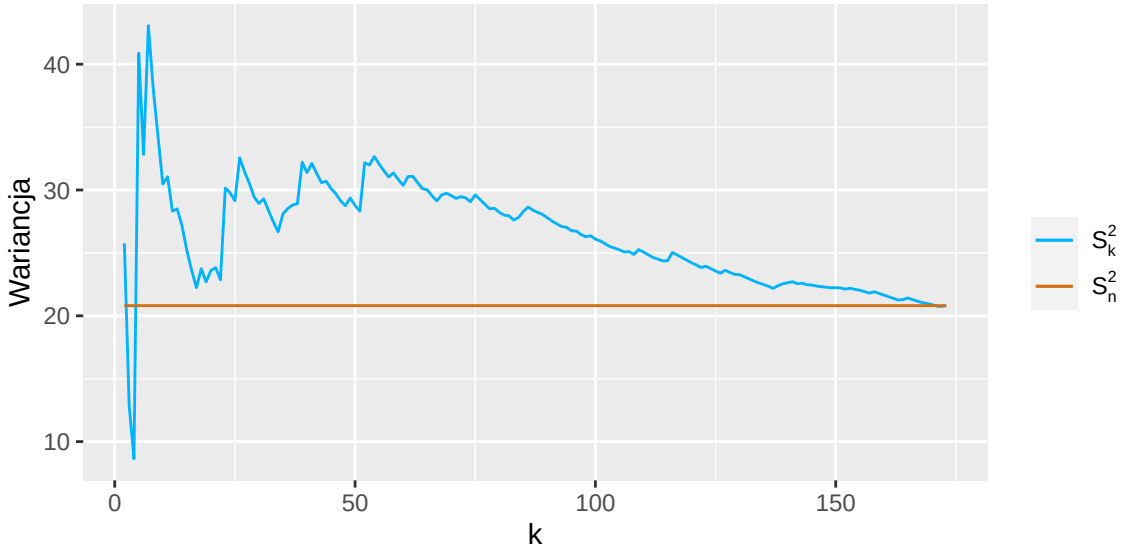
czyli realizacji zmiennych ϵ_i. Ponieważ estymatory $\hat{\beta}_1, \hat{\beta}_0$ zostały wyznaczone metodą najmniejszych kwadratów, to pierwsze założenie na pewno jest spełnione. Dodatkowo, jeśli spojrzymy na wykres 3, residua wydają się być rozłożone losowo wokół zera.



Wykres 3: Wykres punktowy residuów.

W przypadku założenia drugiego obliczymy wariancje częściowe postaci

$$S_k^2 = \frac{1}{k-1}\sum_{i=1}^{k} e_i^2, \quad k = 2, \ldots, n.$$



Wykres 4: Porównanie wariancji częściowej S_k^2 residuów dla $k = 2,..., n$ z wariancją z całej próby, czyli S_n^2.

Na powyższym wykresie zobaczyć możemy, że wariancja częściowa na poczatku ma kilka skoków, jednak wraz ze zwiększającym się k, zaczyna gładko zbiegać do wartości wariancji

z całej próbki. Na tej podstawie możemy zakładać, że $\mathrm{Var}(\epsilon_i)$ jest skończona.

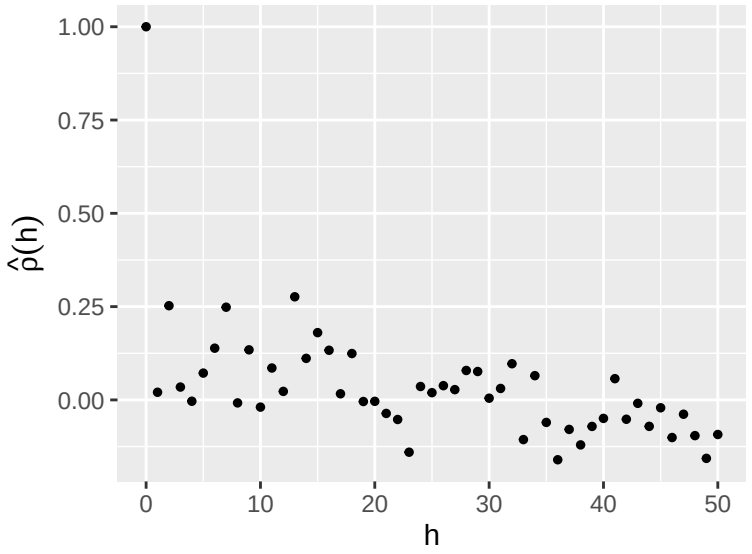
Aby zweryfikować, czy spełnione jest założenie trzecie, skorzystamy z funkcji empirycznej autokorelacji o postaci

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)},$$

gdzie

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{i=1}^{n-|h|} (e_{i+|h|} - \bar{e})(e_i - \bar{e})$$

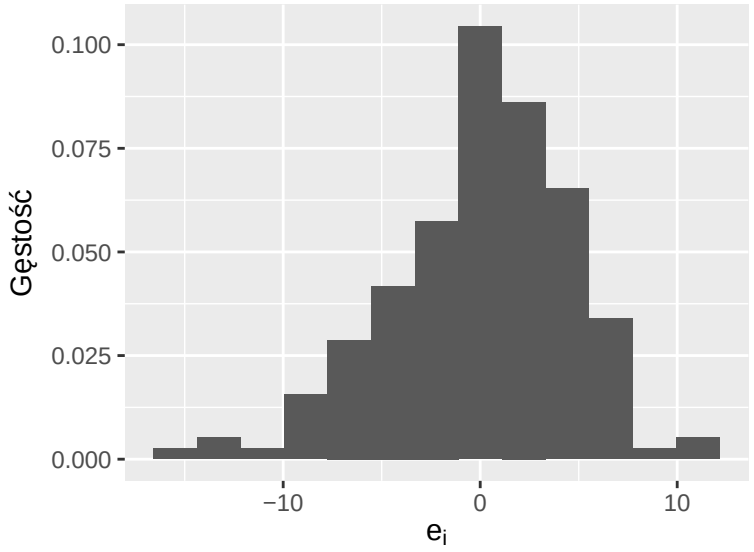
jest estymatorem funkcji autokowariancji.



Wykres 5: Wykres punktowy empirycznej autokorelacji w zależności od przesunięcia h.

Na wykresie 5 widać, że dla $h \neq 0$ funkcja $\hat{\rho}(h)$ oscyluje wokół zera, zatem wnioskujemy, że residua są od siebie niezależne.

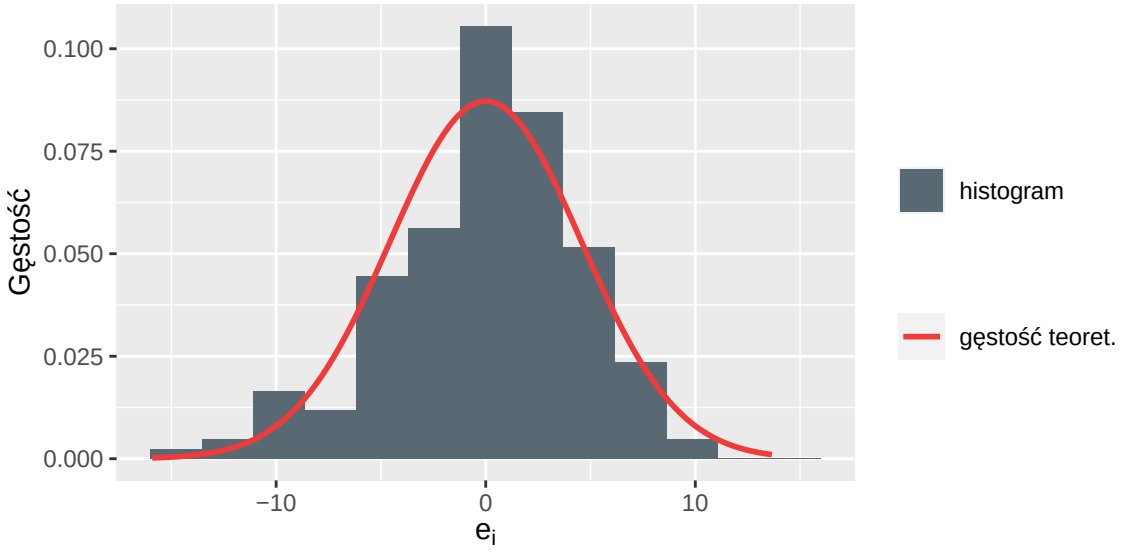
Teraz postaramy się znaleźć rozkład błędów e_i. Zaczniemy od spojrzenia na ich histogram (wykres 6).



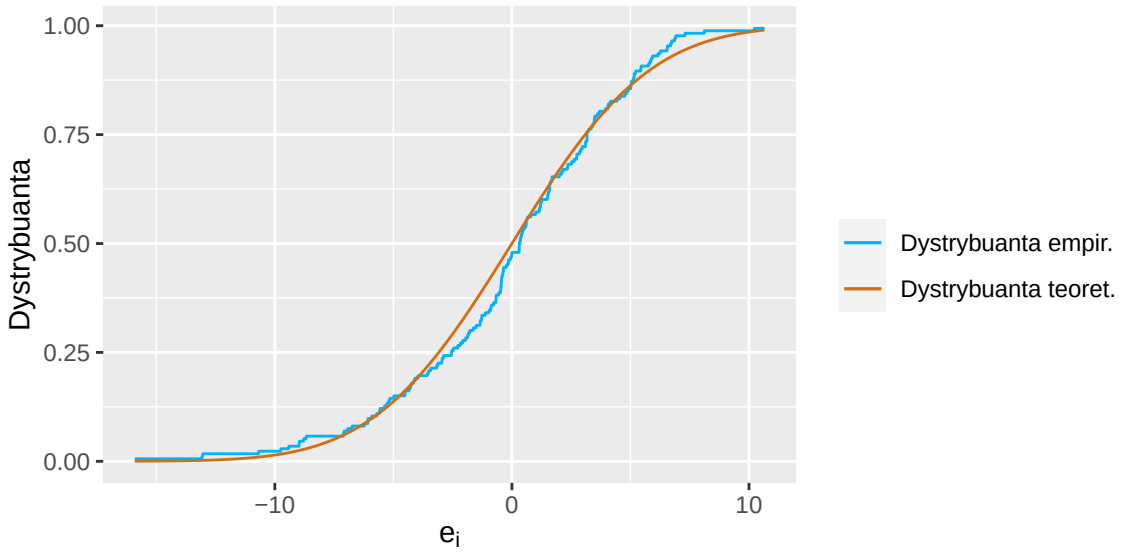
Wykres 6: Histogram residuów.

Kształt histogramu przypomina nieco rozkład normalny. Załóżmy, że w rzeczywistości tak jest i residua pochodzą z rozkładu $\mathcal{N}(0, \sigma^2)$. Wtedy nieobciążonym estymatorem wariancji σ^2 jest

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 \approx 20.93 \ .$$



Wykres 7: Porównanie histogramu z danych z gęstością teoretyczną rozkładu $\mathcal{N}(0, S^2)$.



Wykres 8: Porównanie dystrybuanty empirycznej z danych z dystrybuantą teoretyczną rozkładu $\mathcal{N}(0, S^2)$.

Na wykresie 7 zauważamy, że histogram w miarę pokrywa się z gęstością teoretyczną. Nie spodziewamy się tutaj bardzo dokładnego pokrywania ze względu na niewielką liczbę danych. Z kolei na wykresie 8 widzimy, że dystrybuanta empiryczna wyraźnie nakłada się z dystrybuantą teoretyczną. Aby umocnić nasze przekonania co do normalności residuów, przeprowadzimy test Kołmogorowa-Smirnowa. Przedstawmy hipotezy:

- $\mathcal{H}_0$: wartości residuów są z rozkładu $\mathcal{N}(0, S^2)$
- $\mathcal{H}_1$: wartości residuów nie są z rozkładu $\mathcal{N}(0, S^2)$

Otrzymana p-wartość testu wynosi 0,2532. Ponieważ otrzymany wynik jest wystarczająco duży, to nie mamy podstaw do odrzucenia hipotezy zerowej i możemy przyjąć, że dane pochodzą z rozkładu $\mathcal{N}(0, \sigma^2 = S^2)$.

Przedziały ufności dla β_1 i β_0

Aby skonstruować przedziały ufności dla szukanych parametrów, posłużymy się estymatorami $\hat{\beta}_0$, $\hat{\beta}_1$, ale tym razem pod postacią zmiennych losowych:

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \end{cases} .$$

Zakładając, że $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ można pokazać, że

$$\begin{cases} \hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \ \frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right) \\ \hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \ \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)\right) \end{cases}.$$

Ponieważ nie znamy wartości σ^2, będziemy musieli wykorzystać estymator S^2. Konstruujemy statystyki $\hat{B}_1$ i $\hat{B}_0$ w następujący sposób:

$$\hat{B}_1 = \frac{\hat{\beta}_1 - \beta_1}{S\sqrt{\frac{1}{\sum_{i=1}^n (x_i-\bar{x})^2}}} \sim t_{n-2},$$

$$\hat{B}_0 = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}} \sim t_{n-2},$$

gdzie $S = \sqrt{\frac{\sum_{i=1}^n (Y_i-\hat{Y})^2}{n-2}}$, a t_{n-2} oznacza rozkład t-Studenta z $n-2$ stopniami swobody. Jako pierwszy wyznaczymy przedział ufności dla β_0. Korzystając z powyższych własności, możemy zapisać

$$\mathrm{P}\left(t_{\frac{\alpha}{2},n-2} < \hat{B}_0 < t_{1-\frac{\alpha}{2},n-2}\right) = 1-\alpha,$$

gdzie $t_{\frac{\alpha}{2},n-2}$ i $t_{1-\frac{\alpha}{2},n-2}$ są kwantylami odpowiednio rzędu $\frac{\alpha}{2}$ i $1-\frac{\alpha}{2}$ rozkładu t_{n-2}. Ponieważ rozkład t-Studenta jest symetryczny, to $t_{\frac{\alpha}{2},n-2} = -t_{1-\frac{\alpha}{2},n-2}$. Wykorzystując to oraz podstawiając pod statystykę $\hat{B}_0$ jej wzór, dostajemy

$$\mathrm{P}\left(-t_{1-\frac{\alpha}{2},n-2} < \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}} < t_{1-\frac{\alpha}{2},n-2}\right) = 1-\alpha,$$

co po przekształceniu daje

$$\mathrm{P}\left(\hat{\beta}_0 - t_{1-\frac{\alpha}{2},n-2}S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}} < \beta_0 < \hat{\beta}_0 + t_{1-\frac{\alpha}{2},n-2}S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}\right) = 1-\alpha$$

Zatem przedział ufności dla parametru β_0 na poziomie ufności $1-\alpha$ ma postać

$$\left[\hat{\beta}_0 - t_{1-\frac{\alpha}{2},n-2}S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}} \ , \quad \hat{\beta}_0 + t_{1-\frac{\alpha}{2},n-2}S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}\right].$$

Analogicznie wyznaczamy przedział ufności dla parametru β_1. Wygląda on następująco:

$$\left[\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2} S\sqrt{\frac{1}{\sum_{i=1}^n (x_i-\bar{x})^2}}\ ,\quad \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2} S\sqrt{\frac{1}{\sum_{i=1}^n (x_i-\bar{x})^2}}\ \right].$$

Dla naszych danych przedziały ufności dla parametrów β_0 i β_1 na poziomie ufności 0.95 wynoszą odpowiednio [37.45, 48.35] oraz [1.82, 2.64].

Predykcja

Naszym celem będzie wyznaczenie przedziału ufności dla pewnego

$$Y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0,$$

gdzie $\epsilon_0 \sim \mathcal{N}(0, \sigma^2)$. W tym celu skorzystamy z tego, że

$$\frac{Y_0 - \hat{Y_0}}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - x_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}} \sim t_{n-2},$$

gdzie $\hat{Y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ jest zmienną losową. Możemy zapisać

$$\mathrm{P}\left(-t_{1-\frac{\alpha}{2},n-2} < \frac{Y_0 - \hat{Y_0}}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - x_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}} < t_{1-\frac{\alpha}{2},n-2}\right) = 1 - \alpha,$$

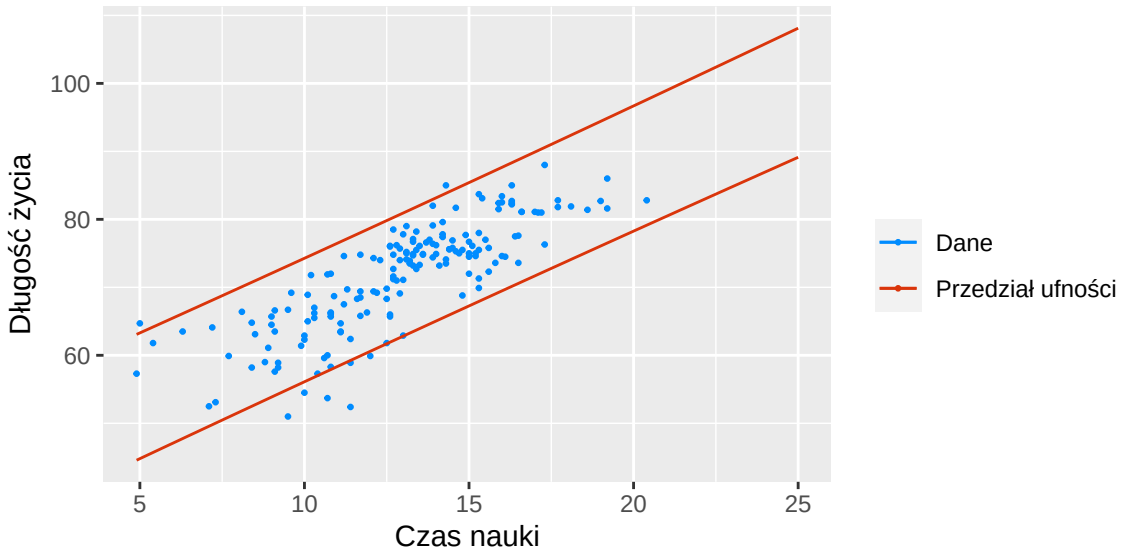
co po przekształceniu prowadzi do

$$\mathrm{P}\left(\hat{Y_0} - t_{1-\frac{\alpha}{2},n-2} S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - x_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}} < Y_0 < \hat{Y_0} + t_{\frac{\alpha}{2},n-2} S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - x_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}\right) = 1-\alpha.$$

Zatem przedział ufności dla Y_0 będzie wyglądał następująco:

$$\left[\hat{Y_0} - t_{1-\frac{\alpha}{2},n-2} S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - x_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}\ ,\ \hat{Y_0} + t_{\frac{\alpha}{2},n-2} S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - x_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}\ \right].$$

Po podstawieniu danych i przyjęciu poziomu ufności 0.95, otrzymaliśmy przedział, który zmienia się w zależności od czasu edukacji (wykres 9). W ten sposób możemy na przykład zobaczyć, w jakim przedziale, z prawdopodobieństwem 0.95, będzie mieściła się długość życia, jeśli średni czas nauki będzie wynosił 25 lat.



Wykres 9: Przedział ufności dla długości życia w zależności od czasu edukacji na poziomie ufności 0.95.

Podsumowanie

Udało nam się wyznaczyć zależność liniową pomiędzy średnia długością życia, a czasem jaki przeciętny obywatel danego państwa poświęca na naukę. Otrzymana przez nas zależność jest dobrze dopasowana, co potwierdza wyliczony współczynnik determinancji równy 0.67. Przy naszych obliczeniach skorzystaliśmy z klasycznego modelu regresji liniowej. Poruszany w raporcie problem długości życia wydaje się być bardziej złożony, gdyż jest on zależny od bardzo wielu różnych czyników. Dobrym pomysłem na kontynuację pracy byłoby znalezienie modelu regresji biorącego pod uwagę inne współczynniki np. poziom opieki medycznej lub wskaźnik GDP danego kraju.