Raport z listy 4

Jakub Kowalczyk

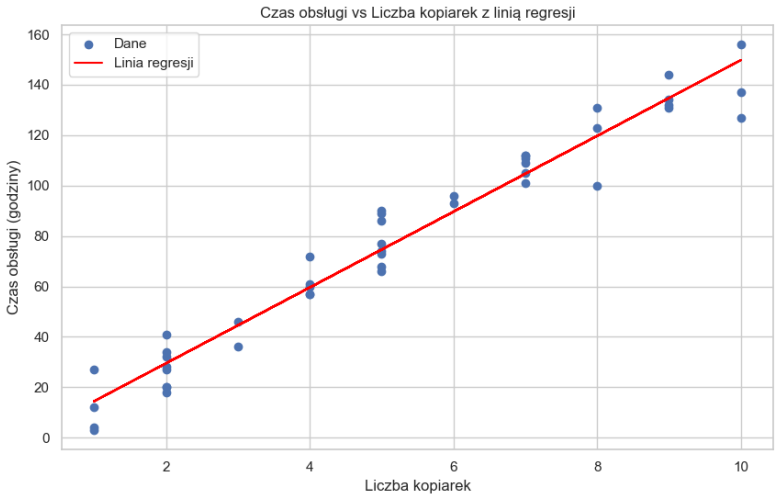
Zadanie 1

W tym zadaniu analizujemy zależność czasu obsługi (y) od liczby obsługiwanych maszyn (x). Dane pochodzą z pliku `ch01pr20.txt`.

a) Wykres i Regresja Liniowa

Na podstawie wykresu punktowego (poniżej) można stwierdzić, że zależność między czasem obsługi a liczbą kopiarek jest w przybliżeniu liniowa. Punkty układają się wzdłuż prostej, co sugeruje zasadność zastosowania modelu regresji liniowej prostej:

$$y = \beta_0 + \beta_1 x + \epsilon$$



Obliczone parametry regresji:

- **Teoretyczne:**
  - $\beta_0 = -0.580157$
  - $\beta_1 = 15.035248$
- **Biblioteczne (`statsmodels`):**
  - $\beta_0 = -0.580157$
  - $\beta_1 = 15.035248$

Równanie regresji:

$$y = -0.58 + 15.04x$$

b) 95% Przedziały Ufności dla Slope'a i Intersepta

Poniższa tabela przedstawia punktowe estymatory oraz 95% przedziały ufności dla wyrazu wolnego (β_0) i współczynnika kierunkowego (β_1). Wyniki obliczone "ręcznie" są zgodne z wynikami z biblioteki.

Parametr	Metryka	Teoretycznie	Biblioteka
Intercept (Beta 0)	Estymator	-0.580157	-0.580157
Slope (Beta 1)	Estymator	15.035248	15.035248
Intercept (Beta 0)	95% CI	(-6.234843, 5.074529)	(-6.234843, 5.074529)
Slope (Beta 1)	95% CI	(14.061010, 16.009486)	(14.061010, 16.009486)

c) Test Istotności Parametrów

Przeprowadzono testy istotności dla obu parametrów. Hipotezy:

- $H_0 : \beta_i = 0$
- $H_1 : \beta_i \neq 0$

Parametr	Metryka	Teoretycznie	Biblioteka
Intercept (Beta 0)	Statystyka T	-0.206908	-0.206908
Intercept (Beta 0)	p-wartość	0.837059	0.837059
Slope (Beta 1)	Statystyka T	31.1233	31.1233
Slope (Beta 1)	p-wartość	0	4.00903e-31

Wnioski:

- **Slope (β_1):** p-wartość $\approx 0 < 0.05$. Odrzucamy H_0. Istnieje istotna statystycznie zależność liniowa między liczbą kopiarek a czasem obsługi. Każda dodatkowa kopiarka zwiększa oczekiwany czas obsługi.
- **Intercept (β_0):** p-wartość $\approx 0.84 > 0.05$. Nie ma podstaw do odrzucenia H_0. Wyraz wolny nie różni się istotnie od zera. W kontekście zadania oznacza to, że przy zerowej liczbie maszyn czas obsługi jest bliski zeru, co jest logiczne.

d) Przedział Ufności dla Wartości Oczekiwanej

Estymujemy średni czas obsługi dla k maszyn.

k	Oczekiwana (Teor)	CI Średnia Dolna (Teor)	CI Średnia Górna (Teor)	Długość CI	Oczekiwana (Bibl)	CI Średnia Dolna (Bibl)	CI Średnia Górna (Bibl)
1	14.4551	9.63614	19.274	9.6379	14.4551	9.63614	19.274
5	74.5961	71.9142	77.2779	5.36372	74.5961	71.9142	77.2779
8	119.702	115.816	123.588	7.77223	119.702	115.816	123.588
11	164.808	158.475	171.14	12.6643	164.808	158.475	171.14
25	375.301	355.74	394.862	39.1219	375.301	355.74	394.862
100	1502.94	1410.46	1595.43	184.966	1502.94	1410.46	1595.43

Wniosek: Długość przedziału ufności zależy od odległości k od średniej liczby maszyn w próbie ($\bar{x} \approx 5.11$). Im dalej punkt k znajduje się od $\bar{x}$, tym przedział jest szerszy (większa niepewność estymacji średniej). Najwęższy przedział obserwujemy dla $k = 5$, które jest najbliżej średniej.

e) Przedział Predykcyjny dla Nowej Obserwacji

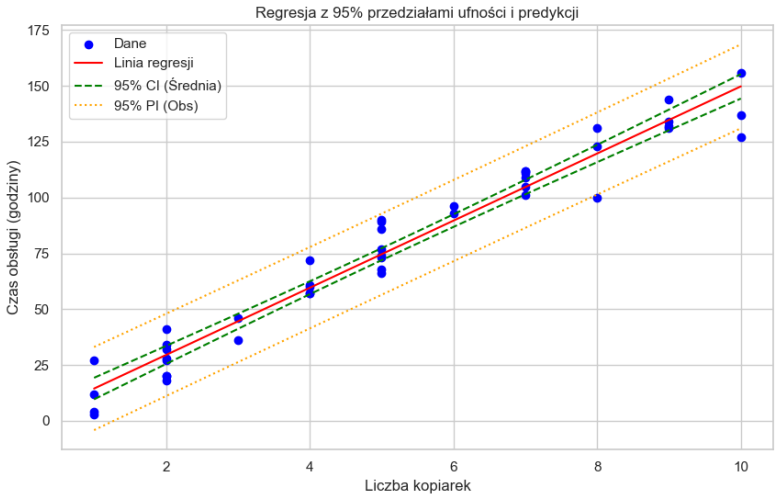
Przewidujemy czas obsługi dla konkretnego przypadku serwisowania k maszyn.

k	Przewidywana (Teor)	PI Dolna (Teor)	PI Górna (Teor)	Długość PI	Przewidywana (Bibl)	PI Dolna (Bibl)	PI Górna (Bibl)
1	14.4551	-4.15544	33.0656	37.2211	14.4551	-4.15544	33.0656
5	74.5961	56.4213	92.7708	36.3495	74.5961	56.4213	92.7708
8	119.702	101.311	138.093	36.7821	119.702	101.311	138.093
11	164.808	145.749	183.866	38.1169	164.808	145.749	183.866
25	375.301	348.735	401.867	53.1323	375.301	348.735	401.867
100	1502.94	1408.73	1597.16	188.428	1502.94	1408.73	1597.16

Wniosek: Podobnie jak w punkcie d), długość przedziału predykcyjnego rośnie wraz z oddalaniem się od średniej. Przedziały te są jednak znacznie szersze niż przedziały ufności dla wartości oczekiwanej.

f) Wizualizacja Przedziałów

Na poniższym wykresie przedstawiono dane, linię regresji oraz 95% przedziały ufności (zielone) i predykcji (pomarańczowe).



Wyjaśnienie różnicy w szerokości przedziałów: Przedziały ufności (CI) dotyczą estymacji **średniej** wartości oczekiwanej ($E[Y|X]$), podczas gdy przedziały predykcyjne (PI) dotyczą **pojedynczej przyszłej obserwacji** (Y_{new}). Przedziały predykcyjne muszą uwzględniać zarówno niepewność co do położenia linii regresji (jak CI), jak i naturalny rozrzut danych wokół tej linii, przez co są zawsze szersze.

Zadanie 2

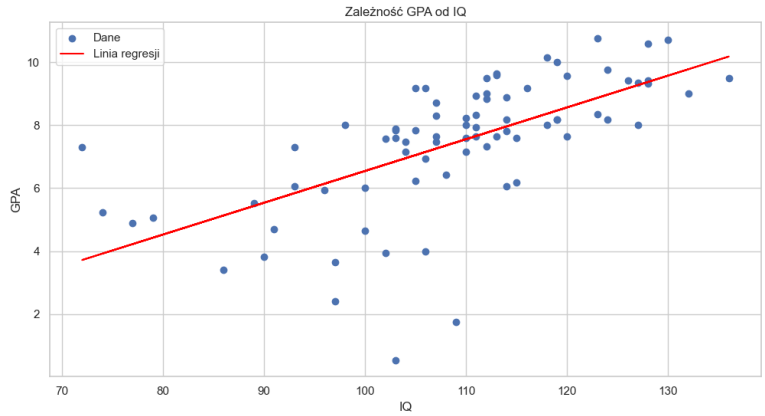
W tym zadaniu analizujemy dane z pliku `tabela1_6.txt` dotyczące 78 uczniów. Badamy zależności między średnią ocen (GPA), wynikiem testu IQ oraz punktacją w teście PH.

a) Regresja liniowa GPA od IQ

Model regresji:

$$GPA = \beta_0 + \beta_1 \cdot IQ + \epsilon$$

Z wykresu (poniżej) widać dodatnią korelację między IQ a GPA.



Równanie regresji:

$$GPA = -3.5571 + 0.1010 \cdot IQ$$

Współczynnik determinacji R^2:

Parametr	Metryka	Teoretycznie	Biblioteka
R^2	Wartość	0.401615	0.401615

Model wyjaśnia około 40.16% zmienności GPA na podstawie wyniku IQ.

b) Test istotności korelacji (Test F)

Testujemy hipotezę, że GPA nie jest skorelowane z IQ, co w modelu liniowym jest równoważne warunkowi, że współczynnik kierunkowy (slope) jest równy 0.

Hipotezy:

- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

Wyniki testu F:

Parametr	Metryka	Teoretycznie	Biblioteka
Statystyka F	Wartość	51.0085	51.0085
P-wartość (F)	Prawdopodobieństwo	4.73734e-10	4.73734e-10

Rozkład F przy założeniu H_0: Rozkład F z $(1, 76)$ stopniami swobody. **Wartość krytyczna** dla $\alpha = 0.05$: $F_{kryt} \approx 3.966760$.

Decyzja: Ponieważ p-wartość ($4.74 \cdot 10^{-10}$) jest mniejsza od poziomu istotności $\alpha = 0.05$ (oraz $F_{stat} > F_{kryt}$), **odrzucamy hipotezę zerową**.

Wniosek: Istnieje istotna statystycznie zależność liniowa między inteligencją (IQ) a średnią ocen (GPA). Wynik testu IQ pozwala w pewnym stopniu przewidywać wyniki w nauce.

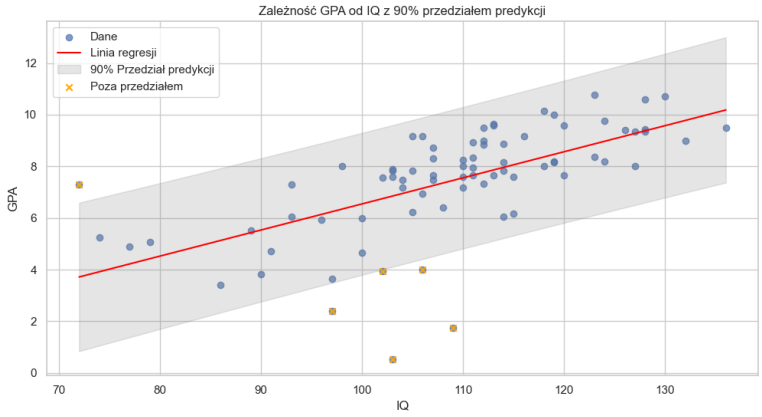
c) Predykcja GPA na podstawie IQ

Przewidywane wartości GPA oraz 90% przedziały predykcyjne dla uczniów o IQ: 75, 100, 140.

IQ	Predicted GPA	Lower 90%	Upper 90%
75.0000	4.0196	1.1659	6.8732
100.0000	6.5451	3.7975	9.2927
140.0000	10.5860	7.7503	13.4216

d) Wykres z 90% przedziałami predykcyjnymi

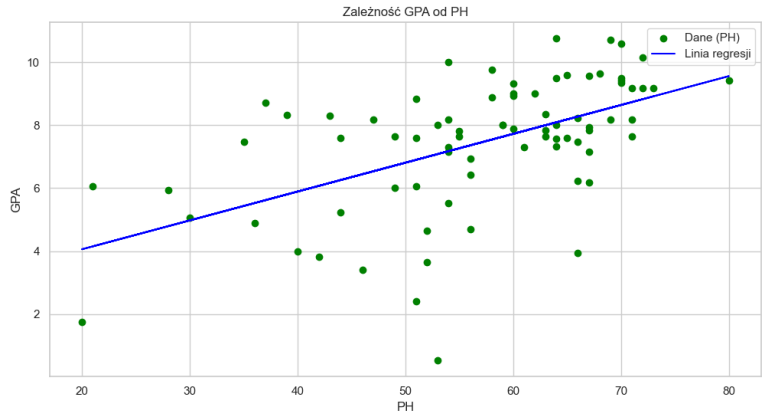
Poniższy wykres przedstawia dane, prostą regresji oraz 90% pasmo predykcji (szary obszar).



Liczba obserwacji poza 90% przedziałem predykcyjnym: **6** (na 78 obserwacji). Stanowi to **7.69%** wszystkich obserwacji, co jest zgodne z oczekiwaniami dla przedziału 90% (oczekujemy ok. 10% obserwacji poza przedziałem).

e) Regresja liniowa GPA od PH

Analiza zależności GPA od wyniku testu PH.



Równanie regresji: $GPA = 2.2259 + 0.0917 * PH$

Porównanie współczynników determinacji R^2:

- Model GPA ~ IQ: $R^2 \approx 0.4016$
- Model GPA ~ PH: $R^2 \approx 0.2936$

Wniosek: Zmienna **IQ** jest lepszym predyktorem średniej ocen (GPA) niż wynik testu PH, ponieważ model oparty na IQ wyjaśnia większą część wariancji GPA (ok. 40% vs 29%).

Zadanie 3

Sprawdzane hipotezy $H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$.

Model generujący dane:

$$Y = 5 + \beta_1 X + \epsilon$$

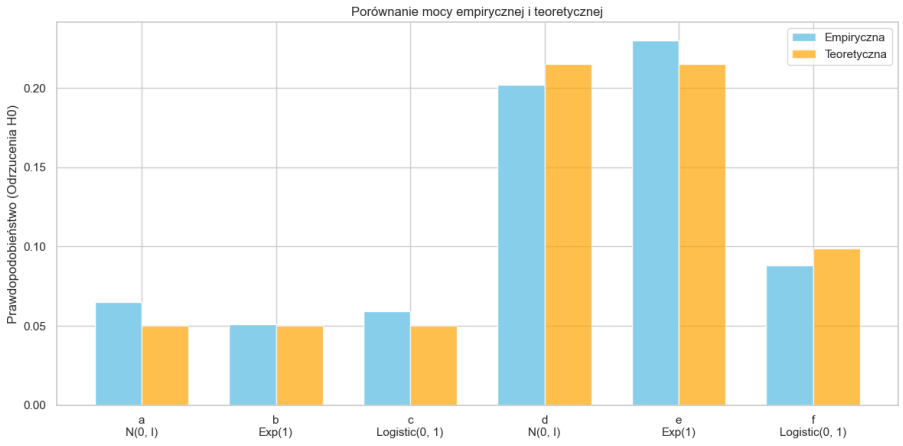
gdzie $X \sim N(0, \frac{1}{500} I)$ jest ustalony dla wszystkich symulacji.

Wyniki Symulacji

Poniższa tabela przedstawia empiryczne prawdopodobieństwo odrzucenia hipotezy H_0 (dla $\alpha = 0.05$) w porównaniu z wartościami teoretycznymi.

Podpunkt	Beta1	Rozkład błędu	Empiryczna	Teoretyczna
a	0	N(0, I)	0.065	0.05 (α)
b	0	Exp(1)	0.056	0.05 (α)
c	0	Logistic(0, 1)	0.042	0.05 (α)
d	2	N(0, I)	0.232	0.215359 (Moc)
e	2	Exp(1)	0.206	0.215359 (Moc)

Podpunkt	Beta1	Rozkład błędu	Empiryczna	Teoretyczna
f	2	Logistic(0, 1)	0.109	0.0987785 (Moc)



Wnioski

1. Błąd I rodzaju (Gdy H_0 jest prawdziwa, $\beta_1 = 0$) W przypadkach a), b) i c), gdzie nie ma rzeczywistej zależności liniowej:

- Oczekiwany poziom błędu I rodzaju wynosi $\alpha = 0.05$.
- Wyniki empiryczne (0.065, 0.056, 0.042) są zbliżone do 0.05.
- **Wniosek:** Test t-Studenta jest **odporny** na naruszenie założenia o normalności rozkładu reszt przy dużej liczbie prób ($n = 200$). Nawet dla niesymetrycznego rozkładu wykładniczego (b) czy grubooogonowego rozkładu logistycznego (c), prawdopodobieństwo błędnego odrzucenia H_0 jest bliskie założonemu poziomowi istotności.

2. Moc testu (Gdy H_0 jest fałszywa, $\beta_1 = 2$) W przypadkach d), e) i f), gdzie istnieje zależność:

- **d) Rozkład Normalny:** Empiryczna moc (0.232) jest zgodna z teoretyczną (0.215). Jest to punkt odniesienia.
- **e) Rozkład Wykładniczy:** Mimo silnej asymetrii rozkładu błędów, moc testu (0.206) jest bardzo zbliżona do mocy dla rozkładu normalnego o tej samej wariancji. Potwierdza to, że przy dużej próbie test t zachowuje swoją moc także dla rozkładów innych niż normalny.
- **f) Rozkład Logistyczny:** Moc testu drastycznie spadła do ok. 0.109. Wynika to z faktu, że standardowy rozkład logistyczny ma wariancję

$\frac{\pi^2}{3} \approx 3.29$, czyli ponad 3-krotnie większą niż standardowy rozkład normalny. Większy szum utrudnia wykrycie zależności.