Laboratorium 5

Wstęp do Analizy Danych | Politechnika Krakowska

Jakub Kapała

Numer albumu: 151885

Data: 31.05.2025

Zadanie 1 - Testowanie hipotez

Treść

1.1 Z biblioteki `MASS` otworzyć zestaw danych `Cars93`. Przedstawić na wykresie pudełkowym średnie ceny samochodów w zależności od producenta. Następnie przedstawić wykresy pudełkowe wyłącznie dla Chevroletów oraz Fordów.
Przydatne funkcje: `subset`, `droplevels`.

1.2 Utworzyć ramki danych zawierające wyłącznie Chevrolety (`Chevrolets93`), Fordy (`Fords93`) oraz jedynie Chevrolety i Fordy (`ChevNFord93`).

1.3 Zbadać hipotezę zerową, że średnia cena Chevroleta na rynku wynosiła $\mu = 15$ (tj. 15 000 $). Obliczyć t-statystykę dla tej hipotezy:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

gdzie $\bar{x}$ – średnia z próbki, μ – średnia zakładana w ramach hipotezy zerowej, s – odchylenie standardowe z próbki, n – rozmiar próbki.

1.4 P-wartość (p-value) to prawdopodobieństwo uzyskania wyników testu co najmniej tak samo skrajnych jak te zaobserwowane w naszej próbce, obliczone przy założeniu, że hipoteza zerowa jest prawdziwa. Należy znaleźć p-wartość dla hipotezy zerowej, że średnia cena Chevroleta na rynku wynosi 15 000 $.

1.5 Znaleźć przedział ufności dla cen Chevroleta odpowiadający poziomowi ufności 95%.

1.6 Znaleźć wielkości wyznaczone w zadaniach 1.3–1.5 za pomocą funkcji `t.test()`.

1.7 Domyślną wartością parametru `alternative` jest hipoteza alternatywna, że prawdziwa wartość średniej różni się od zakładanej w hipotezie zerowej. Przeprowadzić t-test z hipotezą alternatywną, że:

- średnia jest poniżej 15 000 $,
- średnia przekracza 15 000 $.

1.8 Sprawdzić hipotezę zerową, że średnia cena Chevroleta jest równa średniej cenie Forda, z hipotezami alternatywnymi, że cena ta jest:

- różna,
- mniejsza,
- większa.

1.9 Za pomocą funkcji `cor.test()` sprawdzić hipotezę, że rozmiar silnika (`EngineSize`) oraz moc silnika (`Horsepower`) są skorelowane.

Rozwiązanie

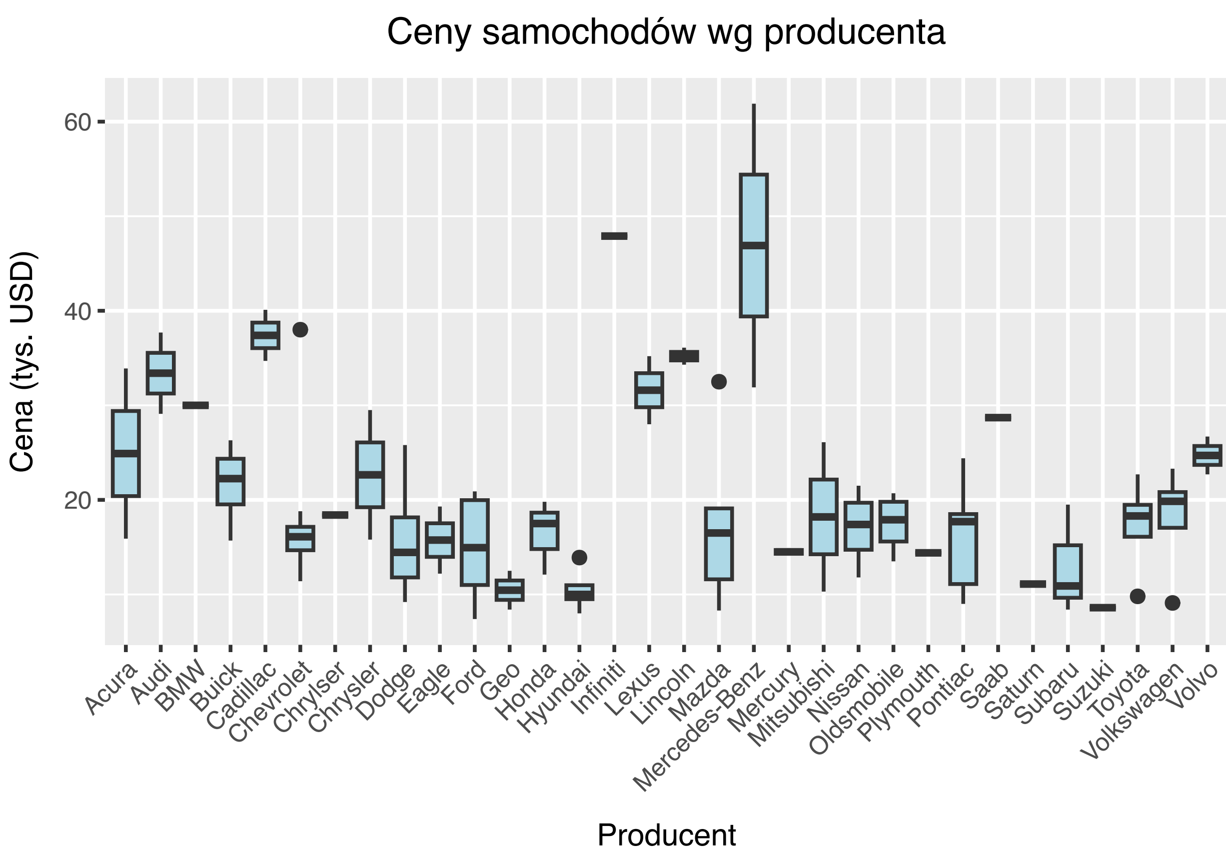
1.1. Utworzenie wykresu średnich cen samochodów w zależności od producenta.

Wczytanie danych:

```
library(MASS)
library(ggplot2)
data(Cars93, package = "MASS")
```

Wyświetlenie wykresu pudełkowego średnich cen samochodów w zależności od producenta:

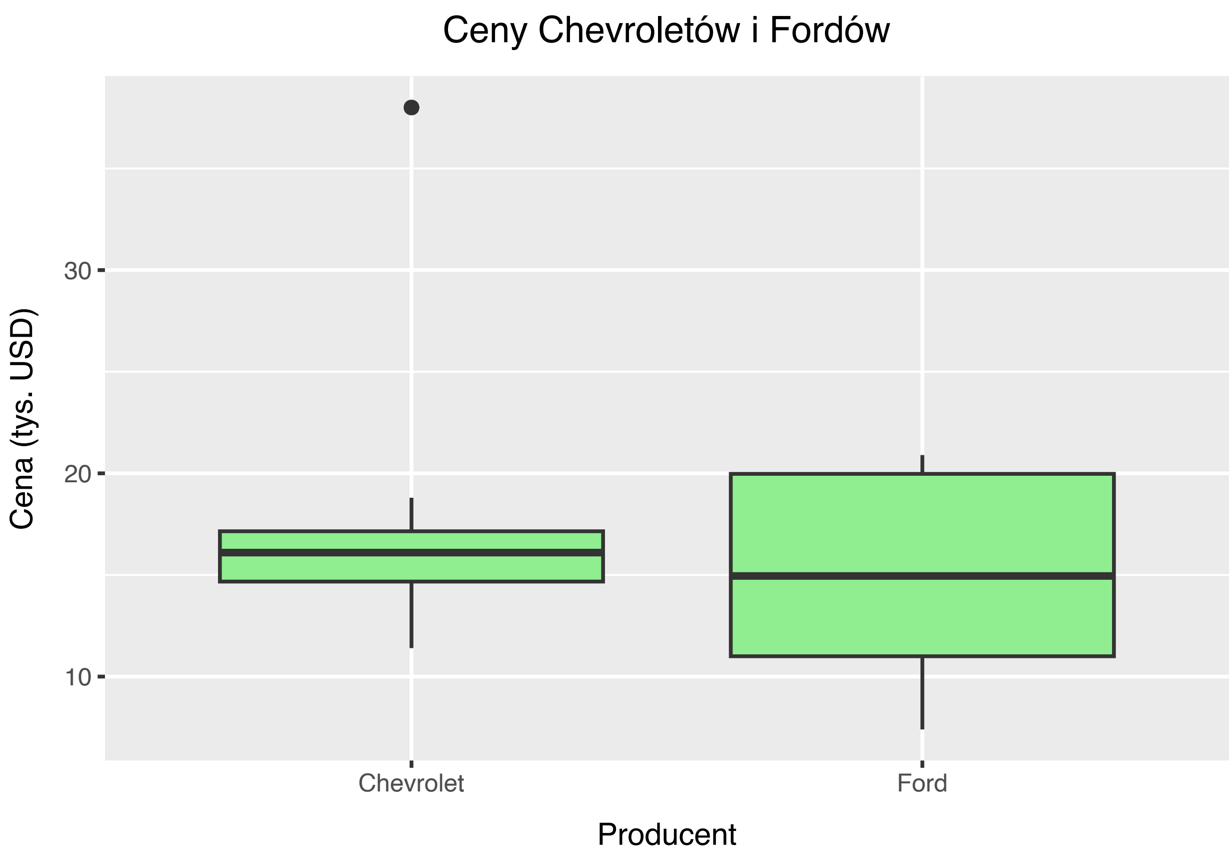
```
ggplot(Cars93, aes(x = Manufacturer, y = Price)) +
  geom_boxplot(fill = "lightblue") +
  labs(
    title = "Ceny samochodów wg producenta",
    x = "Producent", y = "Cena (tys. USD)"
  ) +
  theme(
    plot.title = element_text(hjust = 0.5, margin = margin(b = 10)),
    axis.text.x = element_text(angle = 45, hjust = 1),
    axis.title.y = element_text(margin = margin(r = 10)),
    axis.title.x = element_text(margin = margin(t = 10))
  )
```



Wyświetlenie wykresu pudełkowego tylko dla Chevroletów i Fordów:

```
chev_ford <- droplevels(
  subset(Cars93, Manufacturer %in% c("Chevrolet", "Ford"))
)

ggplot(chev_ford, aes(x = Manufacturer, y = Price)) +
  geom_boxplot(fill = "lightgreen") +
  labs(
    title = "Ceny Chevroletów i Fordów",
    x = "Producent",
    y = "Cena (tys. USD)"
  ) +
  theme(
    plot.title = element_text(hjust = 0.5, margin = margin(b = 10)),
    axis.title.y = element_text(margin = margin(r = 10)),
    axis.title.x = element_text(margin = margin(t = 10))
  )
```



1.2. Utworzenie ramek danych

Ramka danych tylko dla Chevroletów:

```
Chevrolets93 <- droplevels(subset(Cars93, Manufacturer == "Chevrolet"))
```

Ramka danych tylko dla Fordów:

```
Fords93 <- droplevels(subset(Cars93, Manufacturer == "Ford"))
```

Ramka danych dla Chevroletów i Fordów:

```
ChevNFord93 <- droplevels(
  subset(Cars93, Manufacturer %in% c("Chevrolet", "Ford"))
)
```

Wyświetlenie pierwszego wierszu z każdej ramki danych:

```
head(Chevrolets93, 1)
```

Manufacturer Model Type Min.Price Price Max.Price MPG.city MPG.highway 12 Chevrolet Cavalier Compact 8.5 13.4 18.3 25 36 AirBags DriveTrain Cylinders EngineSize Horsepower RPM Rev.per.mile 12 None Front 4 2.2 110 5200 2380 Man.trans.avail Fuel.tank.capacity Passengers Length Wheelbase Width 12 Yes 15.2 5 182 101 66 Turn.circle Rear.seat.room Luggage.room Weight Origin Make 12 38 25 13 2490 USA Chevrolet Cavalier

```
head(Fords93, 1)
```

Manufacturer Model Type Min.Price Price Max.Price MPG.city MPG.highway 31 Ford Festiva Small 6.9 7.4 7.9 31 33 AirBags DriveTrain Cylinders EngineSize Horsepower RPM Rev.per.mile 31 None Front 4 1.3 63 5000 3150 Man.trans.avail Fuel.tank.capacity Passengers Length Wheelbase Width 31 Yes 10 4 141 90 63 Turn.circle Rear.seat.room Luggage.room Weight Origin Make 31 33 26 12 1845 USA Ford Festiva

```
head(ChevNFord93, 1)
```

Manufacturer Model Type Min.Price Price Max.Price MPG.city MPG.highway 12 Chevrolet Cavalier Compact 8.5 13.4 18.3 25 36 AirBags DriveTrain Cylinders EngineSize Horsepower RPM Rev.per.mile 12 None Front 4 2.2 110 5200 2380 Man.trans.avail Fuel.tank.capacity Passengers Length Wheelbase Width 12 Yes 15.2 5 182 101 66 Turn.circle Rear.seat.room Luggage.room Weight Origin Make 12 38 25 13 2490 USA Chevrolet Cavalier

1.3. Badanie hipotezy zerowej

Deklaracja zmiennych - średnia z próbki, średnia z założenia, odchylenie standardowe i rozmiar:

```
mean_price_chevy_0 <- 15
mean_price_chevy_sample <- mean(Chevrolets93$Price)
sd_price_chevy <- sd(Chevrolets93$Price)
n_chevy <- nrow(Chevrolets93)
```

Table 1: Wartości użyte do obliczeń t-statystyki

Średnia z próbki	Średnia z założenia	Odchylenie standardowe	Rozmiar próbki
18.1875	15	8.304463	8

Obliczenie t-statystyki z podanego wzoru:

```
t_stat <- (mean_price_chevy_sample - mean_price_chevy_0) /
  (sd_price_chevy / sqrt(n_chevy))
```

$$t = 1.085634$$

Wynik ten oznacza, że średnia cena Chevroleta w próbie jest oddalona o około 1.09 odchylenia standardowego błędu średniej od średniej zakładanej w hipotezie zerowej. Możemy także dzięki niej obliczyć, jaka dokładnie jest średnia:

```
mean_price_chevy_calc <- t_stat * (sd_price_chevy / sqrt(n_chevy)) + mean_price_chevy_0
mean_price_chevy_calc
```

```
## [1] 18.1875
```

Wartość ta jest spójna z wartością średniej z próby, co potwierdza poprawność obliczeń.

1.4. Znalezienie p-value

Najpierw obliczamy stopnie swobody:

```
df_chevy <- n_chevy - 1
```

Następnie korzystając z t-statystyki z poprzedniego punktu, obliczamy p-value dla hipotezy zerowej, że średnia cena Chevroleta wynosi 15 000 $:

```
p_value_chevy <- 2 * pt(-abs(t_stat), df = df_chevy)
```

$$P = 0.3136$$

Wartość ta jest znacznie większa niż 0.05, co sugeruje, że nie możemy odrzucić hipotezy zerowej. Oznacza to, że średnia cena Chevroleta w próbie nie różni się istotnie od 15 000 $.

1.5. Obliczenie przedziału ufności 95%

Deklaracja poziomu ufności:

```
alpha <- 0.05
```

Obliczanie wartości krytycznej t-Studenta oraz błędu standardowego średniej:

```
t_crit <- qt(1 - alpha/2, df = df_chevy)
se_chevy <- sd_price_chevy / sqrt(n_chevy)
```

Obliczenie przedziału ufności 95% dla średniej ceny Chevroleta:

```
ci_lower <- mean_price_chevy_sample - t_crit * se_chevy
ci_upper <- mean_price_chevy_sample + t_crit * se_chevy

chevy_ci <- c(ci_lower, ci_upper)
```

$$\text{Przedział ufności 95\%:} \quad (11.24,\ 25.13)$$

1.6. Użycie funkcji `t.test()` w celu znalezienia wartości zadań 1.3–1.5

```
nrow(Chevrolets93)
```

```
## [1] 8
```

```
t_test_chevy <- t.test(Chevrolets93$Price, mu = mean_price_chevy_0)
t_test_chevy
```

```
##
##  One Sample t-test
##
## data:  Chevrolets93$Price
## t = 1.0856, df = 7, p-value = 0.3136
## alternative hypothesis: true mean is not equal to 15
## 95 percent confidence interval:
##  11.2448 25.1302
## sample estimates:
## mean of x
##   18.1875
```

$$t = 1.085634$$

$$P = 0.3136069$$

$$\text{Przedział ufności 95\%:} \quad (11.24,\ 25.13)$$

Wartości te są zgodne z obliczeniami wykonanymi w poprzednich punktach.

1.7. Przeprowadzenie t-testu z hipotezą alternatywną, że średnia cena Chevroleta jest poniżej 15 000 $ oraz przekracza 15 000 $.

```
t_test_chevy_below <- t.test(Chevrolets93$Price, mu = mean_price_chevy_0, alternative = "less")
t_test_chevy_above <- t.test(Chevrolets93$Price, mu = mean_price_chevy_0, alternative = "greater")
t_test_chevy_below
```

```
##
##  One Sample t-test
##
## data:  Chevrolets93$Price
## t = 1.0856, df = 7, p-value = 0.8432
## alternative hypothesis: true mean is less than 15
## 95 percent confidence interval:
##      -Inf 23.75012
## sample estimates:
## mean of x
##   18.1875
```

```
t_test_chevy_above
```

```
##
##  One Sample t-test
##
## data:  Chevrolets93$Price
## t = 1.0856, df = 7, p-value = 0.1568
## alternative hypothesis: true mean is greater than 15
## 95 percent confidence interval:
##  12.62488      Inf
## sample estimates:
## mean of x
##   18.1875
```

Table 2: Wyniki t-testów dla hipotez alternatywnych

Hipoteza alternatywna	Przedział ufności 95%	t-statystyka	p-value
Średnia < 15 000	(-Inf, 23.75)	1.085635	0.8431966
Średnia > 15 000	(12.62, Inf)	1.085635	0.1568034

Na bazie wyników t-testów możemy stwierdzić, że

1. dla hipotezy, że średnia cena Chevroleta jest poniżej 15 000 $, p-value jest znacznie większe niż 0.05, co sugeruje, że nie możemy odrzucić hipotezy zerowej. Oznacza to, że nie ma statystycznych dowodów na to, że średnia cena Chevroleta jest niższa niż 15 000 $.

2. dla hipotezy, że średnia cena Chevroleta przekracza 15 000 $, p-value jest większe niż 0.05, co również sugeruje, że nie możemy odrzucić hipotezy zerowej. Oznacza to, że nie ma statystycznych dowodów na to, że średnia cena Chevroleta jest wyższa niż 15 000 $.

Podsumowując, wyniki t-testów potwierdzają, że średnia cena Chevroleta nie różni się istotnie od 15 000 $. Nie oznacza to jednak, że średnia cena Chevroleta jest równa 15 000 $, a jedynie że nie możemy odrzucić hipotez zerowych.

1.8. Sprawdzenie hipotezy zerowej, że średnia cena Chevroleta jest równa średniej cenie Forda.

Najpierw wykonam t-test z hipotezą alternatywną, że średnia cena Chevroleta jest różna od średniej ceny Forda:

```
t_test_chevy_vs_ford_two_sided <- t.test(
  Chevrolets93$Price, Fords93$Price,
  alternative = "two.sided"
)
t_test_chevy_vs_ford_two_sided
```

```
##
##  Welch Two Sample t-test
##
## data:  Chevrolets93$Price and Fords93$Price
## t = 0.93525, df = 11.643, p-value = 0.3687
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.313775 10.763775
## sample estimates:
## mean of x mean of y
##   18.1875   14.9625
```

Następnie sprawdzę hipotezę alternatywną, że średnia cena Chevroleta jest mniejsza od średniej ceny Forda:

```
t_test_chevy_vs_ford_less <- t.test(
  Chevrolets93$Price, Fords93$Price,
  alternative = "less"
)
t_test_chevy_vs_ford_less
```

```
##
##  Welch Two Sample t-test
##
## data:  Chevrolets93$Price and Fords93$Price
## t = 0.93525, df = 11.643, p-value = 0.8157
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 9.386549
## sample estimates:
## mean of x mean of y
##   18.1875   14.9625
```

Ostatnią hipotezą alternatywną będzie, że średnia cena Chevroleta jest większa od średniej ceny Forda:

```
t_test_chevy_vs_ford_greater <- t.test(
  Chevrolets93$Price, Fords93$Price,
  alternative = "greater"
)
t_test_chevy_vs_ford_greater
```

```
##
##  Welch Two Sample t-test
##
## data:  Chevrolets93$Price and Fords93$Price
## t = 0.93525, df = 11.643, p-value = 0.1843
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -2.936549       Inf
## sample estimates:
## mean of x mean of y
##   18.1875   14.9625
```

Table 3: Wyniki t-testów: porównanie średnich cen Chevroletów i Fordów

Hipoteza alternatywna	Przedział ufności 95%	t-statystyka	p-value
Średnie różne	(-4.31, 10.76)	0.9353	0.3687
Chevrolet < Ford	(-Inf, 9.39)	0.9353	0.8157
Chevrolet > Ford	(-2.94, Inf)	0.9353	0.1843

Na podstawie wyników t-testów możemy stwierdzić, że:

1. dla hipotezy, że średnie ceny Chevroletów i Fordów są różne, p-value jest większe od 0.05, co sugeruje, że nie możemy odrzucić hipotezy zerowej. Oznacza to, że nie ma statystycznych dowodów na to, że średnie ceny Chevroletów i Fordów różnią się.
2. dla hipotezy, że średnia cena Chevroleta jest mniejsza od średniej ceny Forda, p-value jest znacznie większe od 0.05, co sugeruje, że nie możemy odrzucić hipotezy zerowej. Oznacza to, że nie ma statystycznych dowodów na to, że średnia cena Chevroleta jest niższa niż średnia cena Forda.
3. dla hipotezy, że średnia cena Chevroleta jest większa od średniej ceny Forda, p-value jest większe od 0.05, co sugeruje, że nie możemy odrzucić hipotezy zerowej. Oznacza to, że nie ma statystycznych dowodów na to, że średnia cena Chevroleta jest wyższa niż średnia cena Forda.

Podsumowując, wyniki t-testów potwierdzają, że średnie ceny Chevroletów i Fordów nie różnią się istotnie. Nie oznacza to jednak, że średnie ceny są równe, a jedynie że nie możemy odrzucić hipotez zerowych.

1.9. Sprawdzenie hipotezy, że rozmiar silnika (`EngineSize`) oraz moc silnika (`Horsepower`) są skorelowane.

```
cor_test_engine_horsepower <- cor.test(
  Cars93$EngineSize, Cars93$Horsepower,
  method = "pearson"
)
cor_test_engine_horsepower
```

```
##
##  Pearson's product-moment correlation
##
## data:  Cars93$EngineSize and Cars93$Horsepower
## t = 10.253, df = 91, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6210314 0.8143733
## sample estimates:
##       cor
## 0.7321197
```

Table 4: Wyniki testu korelacji: rozmiar silnika vs moc silnika

	Współczynnik korelacji	Przedział ufności 95%	p-value
cor	0.7321	(0.621, 0.8144)	0

Na podstawie testu korelacji Pearsona możemy stwierdzić, że:

1. Współczynnik korelacji pomiędzy rozmiarem silnika a mocą silnika wynosi $r = 0.73$, co wskazuje na silną dodatnią zależność liniową.
2. Przedział ufności 95% dla współczynnika korelacji to $(0.62,\ 0.81)$.
3. p-value jest znacznie mniejsze niż 0.05, co pozwala odrzucić hipotezę zerową o braku korelacji.

Podsumowując, istnieje istotna statystycznie, silna dodatnia korelacja pomiędzy rozmiarem silnika a mocą silnika w zbiorze Cars93.

Zadanie 2

Treść Utworzyć funkcje `slope(x, y)` oraz `intercept(x, y)`, które otrzymawszy wektory współrzędnych x i y danych zwracają współczynnik kierunkowy a oraz wyraz wolny b dopasowanej do tych danych funkcji liniowej zgodnie ze wzorami metody najmniejszych kwadratów:

$$a = \frac{nS_{xy} - S_x S_y}{nS_{xx} - S_x^2} \qquad b = \frac{S_y S_{xx} - S_x S_{xy}}{nS_{xx} - S_x^2},$$

gdzie:

$$S_x = \sum_i x_i, \quad S_y = \sum_i y_i, \quad S_{xx} = \sum_i x_i^2, \quad S_{xy} = \sum_i x_i y_i.$$

Rozwiązanie Funkcja slope oblicza współczynnik kierunkowy a:

```r
slope <- function(x, y) {
  n <- length(x)
  Sx <- sum(x)
  Sy <- sum(y)
  Sxx <- sum(x^2)
  Sxy <- sum(x * y)
  a <- (n * Sxy - Sx * Sy) / (n * Sxx - Sx^2)
  return(a)
}
```

Funkcja intercept oblicza wyraz wolny b:

```r
intercept <- function(x, y) {
  n <- length(x)
  Sx <- sum(x)
  Sy <- sum(y)
  Sxx <- sum(x^2)
  Sxy <- sum(x * y)
  b <- (Sy * Sxx - Sx * Sxy) / (n * Sxx - Sx^2)
  return(b)
}
```

Test działania funkcji:

```r
x <- c(1, 2, 3, 4, 5)
y <- c(2, 3, 5, 7, 11)
a <- slope(x, y)
b <- intercept(x, y)
cat("Równanie prostej: y =", a, "* x +", b, "\n")
```

```
## Równanie prostej: y = 2.2 * x + -1
```

Porównanie do wbudowanych funkcji:

```r
lm_model <- lm(y ~ x)
cat("Równanie prostej z lm(): y =", coef(lm_model)[2], "* x +", coef(lm_model)[1], "\n")
```

```
## Równanie prostej z lm(): y = 2.2 * x + -1
```

Zadanie 3

Treść

3.1 Z biblioteki `MASS` otworzyć zestaw danych `crabs` dotyczący krabów z gatunku *Leptograpsus variegatus*[1]. Przefiltrować dane tak, by zawierały wyłącznie samców z gatunku niebieskiego (B). Przeprowadzić regresję liniową dla zależności długości pancerza (CL) od szerokości pancerza (CW) tych osobników. Przedstawić wykres tej zależności wraz z dopasowaną prostą.
Potrzebne funkcje: `lm()`, `summary()`, `abline()`.

3.2 Znaleźć współczynniki prostej dopasowanej za pomocą metody najmniejszych kwadratów, korzystając z funkcji z zadania 2.

3.3 Przedstawić na wykresie błędy (residuals) między długością pancerza przewidzianą na podstawie dopasowanej prostej a rzeczywistą.

Rozwiązanie

3.1. Wczytanie danych i przefiltrowanie ich

Wczytanie danych:

```r
library(MASS)
data(crabs, package = "MASS")
```

Przefiltrowanie danych, aby zawierały jedynie samce z gatunku niebieskiego (`B`):

```r
crabs_male_blue <- subset(crabs, sex == "M" & sp == "B")
```

Regresja liniowa dla zależności długości pancerza (CL) od szerokości pancerza (CW):

```r
lm_crabs <- lm(CL ~ CW, data = crabs_male_blue)
```

Podsumowanie modelu regresji:

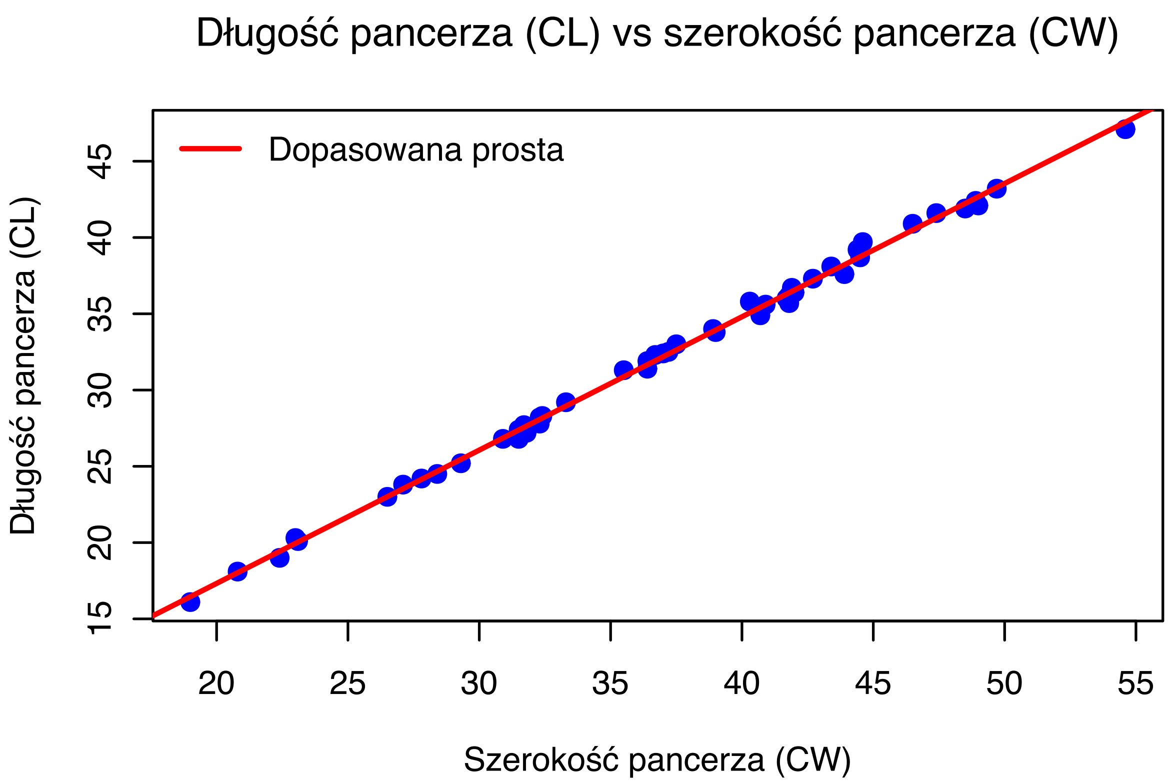
```r
summary(lm_crabs)
```

```
##
## Call:
## lm(formula = CL ~ CW, data = crabs_male_blue)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67482 -0.25450  0.01909  0.24268  0.87823
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.154661   0.229214  -0.675    0.503
## CW           0.873911   0.006076 143.841   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3553 on 48 degrees of freedom
## Multiple R-squared:  0.9977, Adjusted R-squared:  0.9976
## F-statistic: 2.069e+04 on 1 and 48 DF,  p-value: < 2.2e-16
```

[1] https://en.wikipedia.org/wiki/Leptograpsus

Teraz wyświetlę wykres zależności długości pancerza od szerokości pancerza wraz z dopasowaną prostą:

```
plot(crabs_male_blue$CW, crabs_male_blue$CL,
     main = "Długość pancerza (CL) vs szerokość pancerza (CW)",
     xlab = "Szerokość pancerza (CW)",
     ylab = "Długość pancerza (CL)",
     pch = 19, col = "blue")
abline(lm_crabs, col = "red", lwd = 2)
legend("topleft", legend = "Dopasowana prosta", col = "red", lwd = 2, bty = "n")
```



3.2. Współczynniki prostej dopasowanej za pomocą metody najmniejszych kwadratów

Współczynnik kierunkowy (nachylenie) i wyraz wolny prostej dopasowanej do danych:

```
a_crabs <- slope(crabs_male_blue$CW, crabs_male_blue$CL)
b_crabs <- intercept(crabs_male_blue$CW, crabs_male_blue$CL)
```

$$\mathrm{a} = 0.8739$$

$$\mathrm{b} = -0.1547$$

$$\text{Równanie prostej: } y = 0.8739 \cdot x - 0.1547$$

Wartości te są zgodne z wynikami uzyskanymi w poprzednim punkcie z funkcji `lm()`.

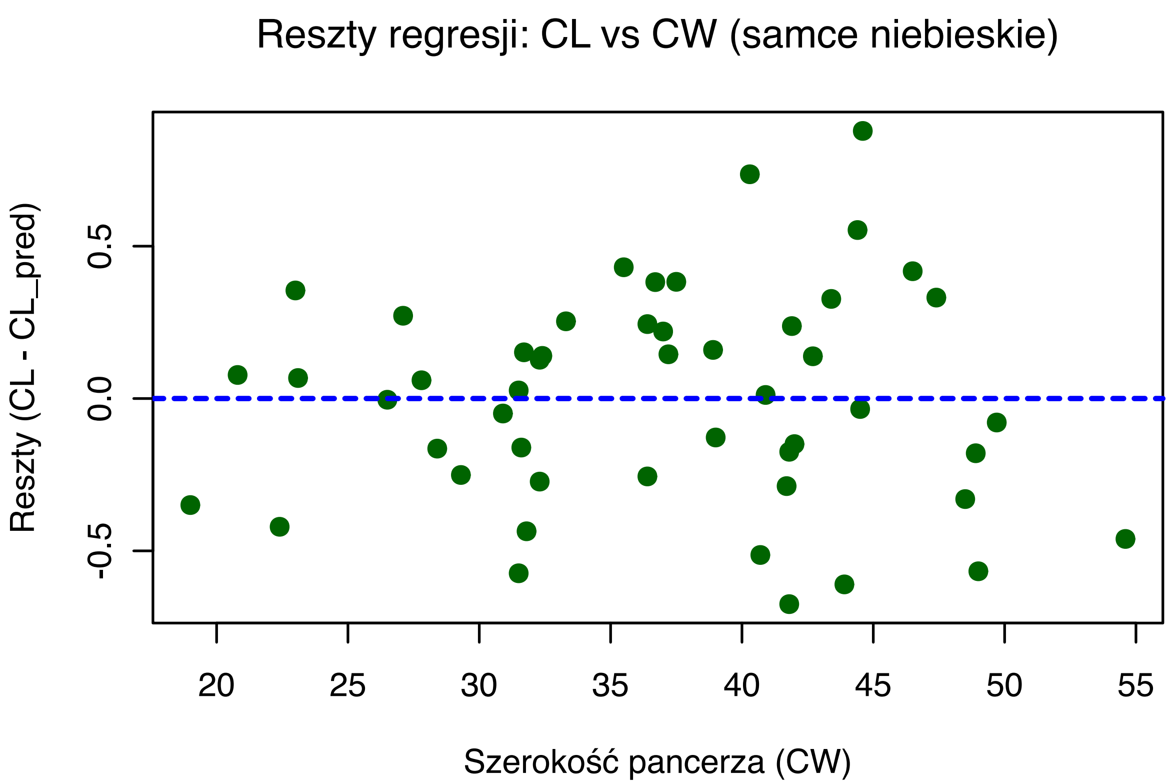
3.3. Wykres błędów (residuals)

Obliczenie błędów (residuals) między długością pancerza przewidzianą na podstawie dopasowanej prostej a rzeczywistą:

```
predicted_cl <- a_crabs * crabs_male_blue$CW + b_crabs
residuals_crabs <- crabs_male_blue$CL - predicted_cl
```

Wyświetlanie wykresu:

```
plot(crabs_male_blue$CW, residuals_crabs,
     main = "Reszty regresji: CL vs CW (samce niebieskie)",
     xlab = "Szerokość pancerza (CW)",
     ylab = "Reszty (CL - CL_pred)",
     pch = 19, col = "darkgreen")
abline(h = 0, col = "blue", lwd = 2, lty = 2)
```



Zadanie 4

Treść

4.1 Z biblioteki `MASS` otworzyć zestaw danych `steam` dotyczący zależności ciśnienia pary nasyconej od temperatury. Przeprowadzić regresję liniową dla zależności $p(T)$.

4.2 Przedstawić na wykresie zależność ciśnienia pary nasyconej od temperatury wraz z dopasowaną prostą.

4.3 Przedstawić wykres błędów dopasowania prostej (residuals).

4.4 Powyższe dane są słabo opisywane przez funkcję liniową. Spróbować dopasować funkcję kwadratową do danych. Przedstawić wykres punktów pomiarowych wraz z dopasowaną krzywą oraz wykres błędów.

Rozwiązanie

4.1. Wczytanie danych i przeprowadzenie regresji liniowej

Wczytanie danych:

```r
library(MASS)
data(steam, package = "MASS")
```

Przeprowadzenie regresji liniowej dla zależności ciśnienia pary nasyconej od temperatury:

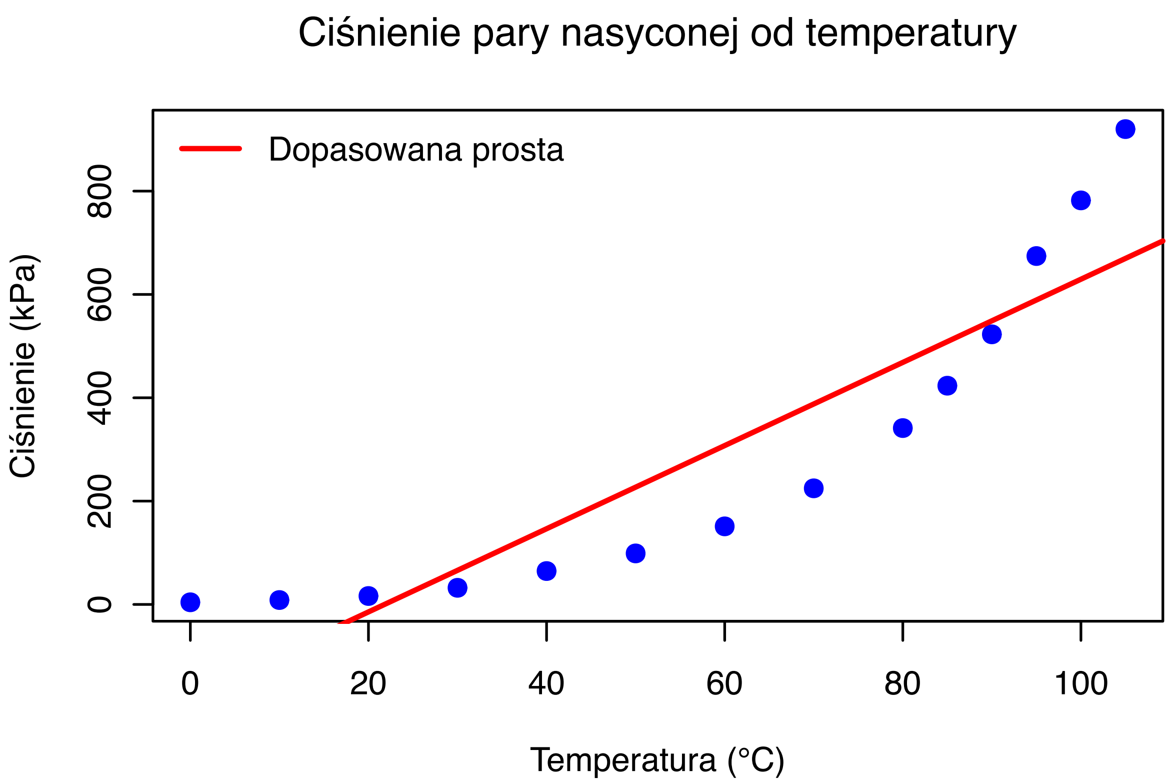
```r
lm_steam <- lm(Press ~ Temp, data = steam)
summary(lm_steam)
```

```
##
## Call:
## lm(formula = Press ~ Temp, data = steam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -163.22 -116.66  -29.98   98.93  250.32
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -175.503     76.156  -2.305   0.0399 *
## Temp           8.049      1.111   7.245 1.02e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 140.4 on 12 degrees of freedom
## Multiple R-squared:  0.8139, Adjusted R-squared:  0.7984
## F-statistic: 52.49 on 1 and 12 DF,  p-value: 1.022e-05
```

4.2. Wykres zależności ciśnienia pary nasyconej od temperatury wraz z dopasowaną prostą

Wyświetlenie wykresu zależności ciśnienia pary nasyconej od temperatury wraz z dopasowaną prostą:

```
plot(steam$Temp, steam$Press,
     main = "Ciśnienie pary nasyconej od temperatury",
     xlab = "Temperatura (°C)",
     ylab = "Ciśnienie (kPa)",
     pch = 19, col = "blue")
abline(lm_steam, col = "red", lwd = 2)
legend("topleft", legend = "Dopasowana prosta", col = "red", lwd = 2, bty = "n")
```



Model regresji liniowej jest widoczny na wykresie jako czerwona linia. Widać, że model ten nie jest idealnym dopasowaniem do danych, ponieważ punkty pomiarowe nie leżą blisko linii regresji. Lepsza byłaby funkcja nieliniowa, np. kwadratowa.

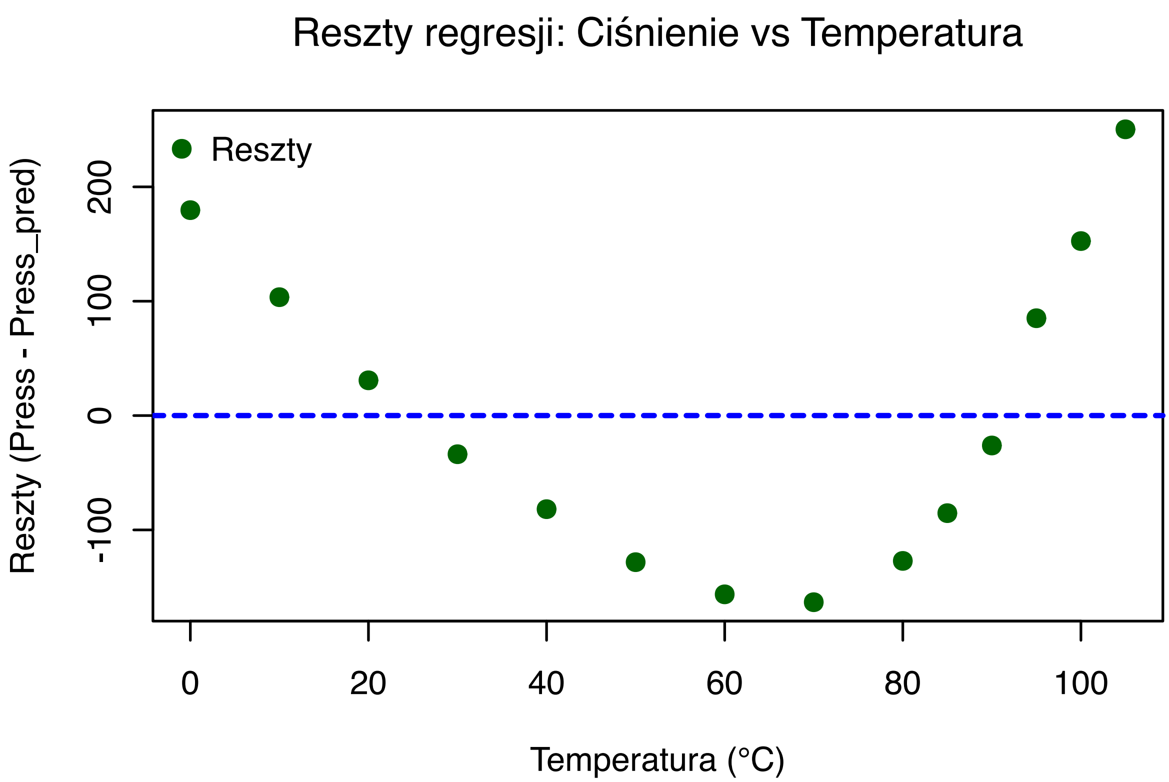
4.3. Wykres błędów dopasowania prostej (residuals)

Obliczenie błędów (residuals) między ciśnieniem przewidzianym na podstawie dopasowanej prostej a rzeczywistym:

```
predicted_press <- predict(lm_steam)
residuals_steam <- steam$Press - predicted_press
```

Wyświetlenie wykresu błędów:

```
plot(steam$Temp, residuals_steam,
     main = "Reszty regresji: Ciśnienie vs Temperatura",
     xlab = "Temperatura (°C)",
     ylab = "Reszty (Press - Press_pred)",
     pch = 19, col = "darkgreen")
abline(h = 0, col = "blue", lwd = 2, lty = 2)
legend("topleft", legend = "Reszty", col = "darkgreen", pch = 19, bty = "n")
```



Jak widać na wykresie, reszty znacznie różnią się od zera, co sugeruje, że model liniowy nie jest najlepszym dopasowaniem do danych.

4.4. Dopasowanie funkcji kwadratowej do danych

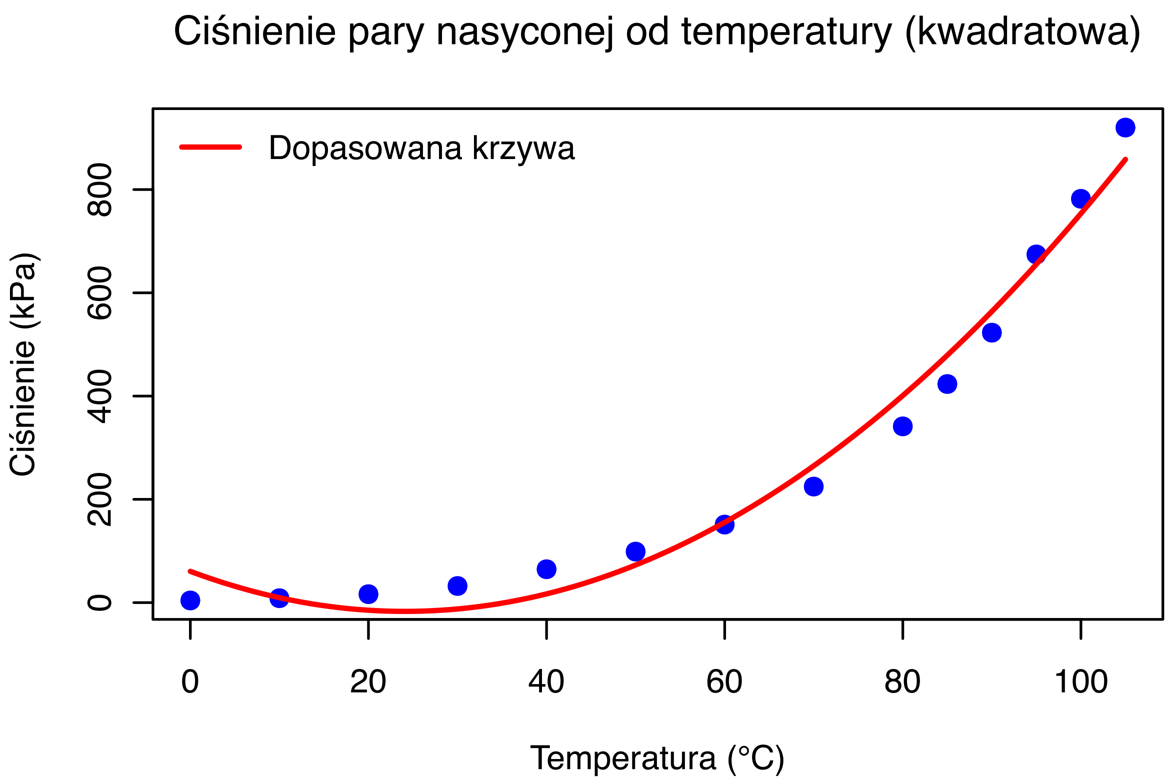
Przeprowadzenie regresji kwadratowej:

```
lm_steam_quad <- lm(Press ~ poly(Temp, 2, raw = TRUE), data = steam)
summary(lm_steam_quad)
```

```
##
## Call:
## lm(formula = Press ~ poly(Temp, 2, raw = TRUE), data = steam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.670 -40.871   8.984  30.440  61.675
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                60.53091   34.86869   1.736  0.11046
## poly(Temp, 2, raw = TRUE)1 -6.43837    1.51500  -4.250  0.00137 **
## poly(Temp, 2, raw = TRUE)2  0.13368    0.01356   9.861  8.5e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.75 on 11 degrees of freedom
## Multiple R-squared:  0.9811, Adjusted R-squared:  0.9777
## F-statistic: 285.4 on 2 and 11 DF,  p-value: 3.324e-10
```

Wyświetlenie wykresu punktów pomiarowych wraz z dopasowaną krzywą:

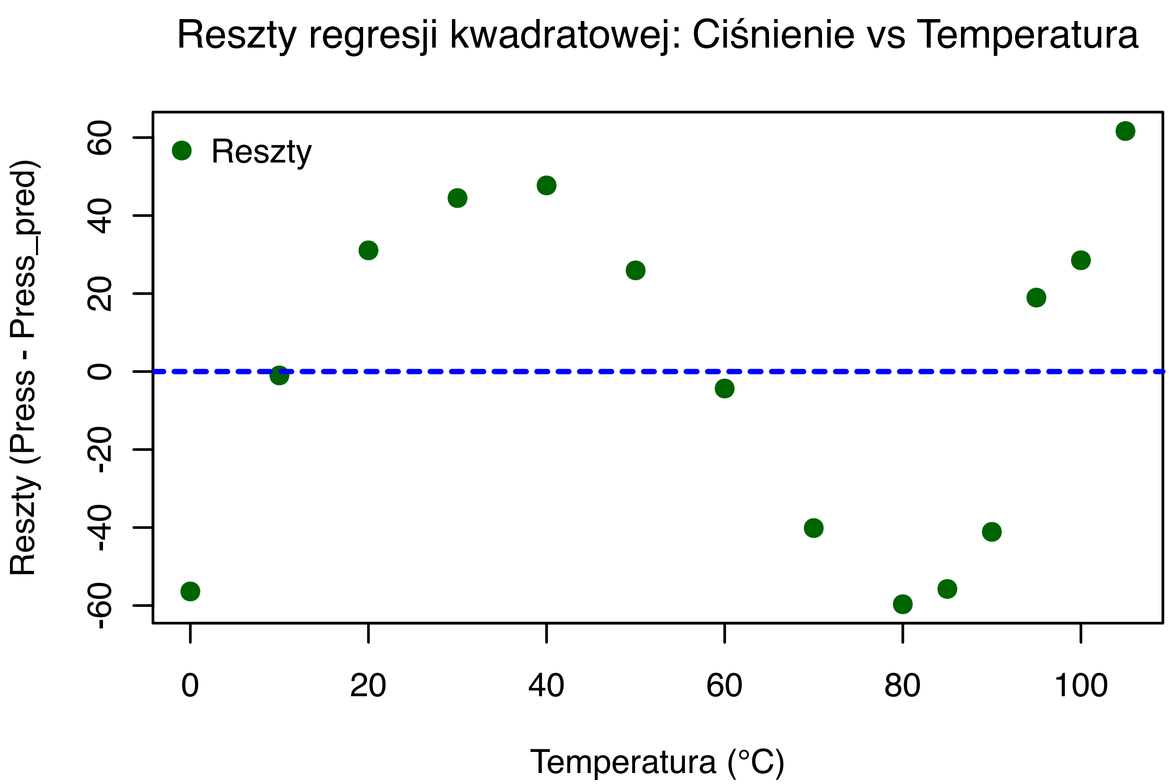
```
plot(steam$Temp, steam$Press,
     main = "Ciśnienie pary nasyconej od temperatury (kwadratowa)",
     xlab = "Temperatura (°C)",
     ylab = "Ciśnienie (kPa)",
     pch = 19, col = "blue")
curve(predict(lm_steam_quad, newdata = data.frame(Temp = x)),
      add = TRUE, col = "red", lwd = 2)
legend("topleft", legend = "Dopasowana krzywa", col = "red", lwd = 2, bty = "n")
```



Tym razem wygląda to znacznie lepiej. Krzywa dopasowana do danych jest bardziej złożona i lepiej odwzorowuje zależność między temperaturą a ciśnieniem pary nasyconej.

Zobaczmy teraz wykres błędów dopasowania tej krzywej:

```
predicted_press_quad <- predict(lm_steam_quad)
residuals_steam_quad <- steam$Press - predicted_press_quad
plot(steam$Temp, residuals_steam_quad,
     main = "Reszty regresji kwadratowej: Ciśnienie vs Temperatura",
     xlab = "Temperatura (°C)",
     ylab = "Reszty (Press - Press_pred)",
     pch = 19, col = "darkgreen")
abline(h = 0, col = "blue", lwd = 2, lty = 2)
legend("topleft", legend = "Reszty", col = "darkgreen", pch = 19, bty = "n")
```



Na wykresie błędów widać, że reszty są znacznie mniejsze niż w przypadku modelu liniowego, co sugeruje, że model kwadratowy lepiej dopasowuje się do danych.