Iga Świtalska, Karolina Rakus

Regresja liniowa na podstawie danych rzeczywistych

5 lutego 2022

1. Opis danych

Dane, które będziemy analizować dotyczą cech fizycznych krabów. Parametry te zostały zmierzone w okolicach Bostonu. Dane zawierają następujące kolumny wymienione poniżej, każda z nich zawiera 3893 obserwacje (nie ma żadnych braków danych). Kolumnę "Viscera Weight" zdecydowałyśmy się usunąć ze względu na trudności z jednoznaczną jej interpretacją.

- Sex - płeć (zmienna kategoryczna),
  - M - male (samiec),
  - F - female (samica),
  - I - indeterminate (nieokreślona),
- Length - długość kraba mierzona w stopach (zmienna ciągła),
- Diameter - średnica kraba mierzona w stopach (zmienna numeryczna),
- Height - wysokość kraba mierzona w stopach (zmienna numeryczna),
- Weight - waga kraba mierzona w uncjach (zmienna numeryczna),
- Schuncked Weight - waga kraba bez muszli mierzona w uncjach (zmienna numeryczna),
- Shell Weight - waga muszli mierzona w uncjach (zmienna numeryczna),
- Age - wiek kraba w miesiącach (zmienna numeryczna).

Pierwsze 6 wierszy naszego zbioru danych można zobaczyć w tabeli 1.1.

	Sex	Length	Diameter	Height	Weight	Shucked.Weight	Shell.Weight	Age
1	F	1.44	1.18	0.41	24.64	12.33	6.75	9
2	M	0.89	0.65	0.21	5.40	2.30	1.56	6
3	I	1.04	0.78	0.25	7.95	3.23	2.76	6
4	F	1.18	0.89	0.25	13.48	4.75	5.24	10
5	I	0.89	0.66	0.21	6.90	3.46	1.70	6
6	F	1.55	1.16	0.35	28.66	13.58	7.23	8

Tabela 1.1. Pierwsze 6 obserwacji

Do dalszej analizy wybrałyśmy dwie zmienne: wysokość ("Height") - zmienna obiaśniana i średnicę ("Diameter") - zmienna objasniająca. Spróbujemy w dalszej części analizy zaimplementować dla tych zmiennych model regresji liniowej.

2. Analiza zmiennej objaśnianej

	Height
X	Min. :0.0000
X.1	1st Qu.:0.2875
X.2	Median :0.3625
X.3	Mean :0.3494
X.4	3rd Qu.:0.4125
X.5	Max. :2.8250
X.6	Variance: 0.011
X.7	Skewness: 3.3131
X.8	Kurtosis: 83.149
X.9	Variability: 0.3005

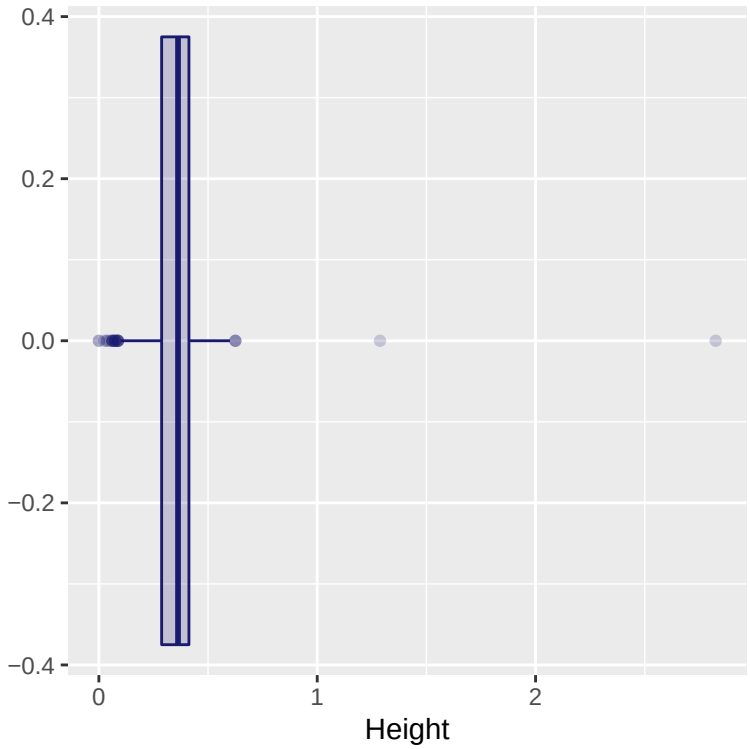
Tabela 2.1. Statystyki opisowe zmiennej objaśnianej przed ususnięciem obserwacji odstających

Na wykresie pudełkowym 2.1 zauważamy dwie odstające obserwacje. Odczytując dokładne ich wartości otrzymujemy: 2.825 stóp = 86.106 cm, 1.2875 stóp = 39.243 cm. Możliwą przyczyną odstających danych mógł być błąd pomiaru. Niestety nie jesteśmy w stanie tego spradzić. Są to jednak wartości na tyle rzadkie w naszym zbiorze danych, że zdecydowałyśmy się je usunąć. Mogłyby bowiem mieć wpływ na późniejszy model regresji liniowej.

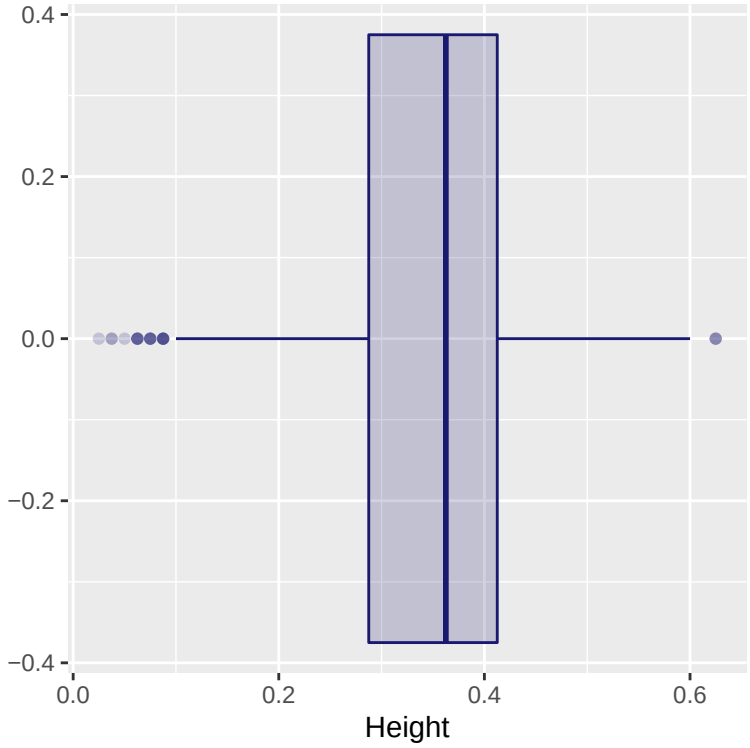
W tabeli 2.1 widać również, że najmniejszą wartością jaką przyjmuje wysokość jest 0. Nie jest to realistyczne i po raz kolejny mogło być wynikiem błędu pomiarowego. Dlatego postanowiłyśmy również usunąć wszystkie obserwacje, które mają wysokość równą 0. Dalsza część analiz jest przeprowadzona dla zmodyfikowanych danych.

Zanim jednak usuniemy wartości odstające srawdzimy ile wynosi wariancja i jak zmieni się ona po usunięciu danych. Wariancja przed usunięciem wartości odstających wynosi:0.01102.

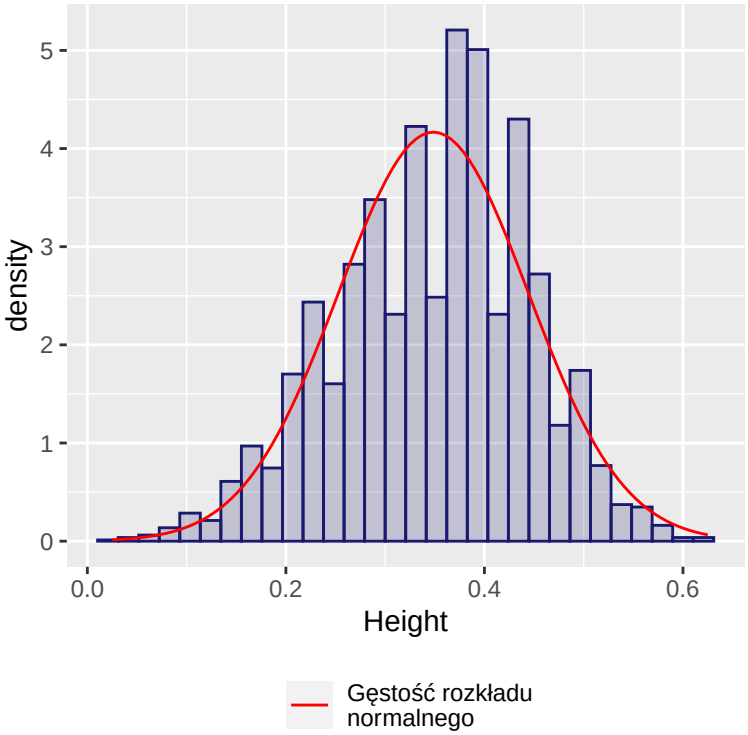
Po usunięciu: 0.0091654. Widać, że wariancja zmniejszyła się, podobnie jak średnia, zmienność i kurtoza. Na wykresie pudełkowym 2.2 również możemy zobaczyć różnicę. Po usunięciu obserwacji odstających i nietypowych, dane mają rozkład dużo bardziej symetryczny. Pozostałe statystki opisowe takie jak kwartyle i mediana nie zmieniły się, ponieważ nie są czułe na obserwacje odstające. Wartości statystyk przed i po usunięciu skrajnych danych można porównać, patrząc na tabele 2.1 i 2.2.



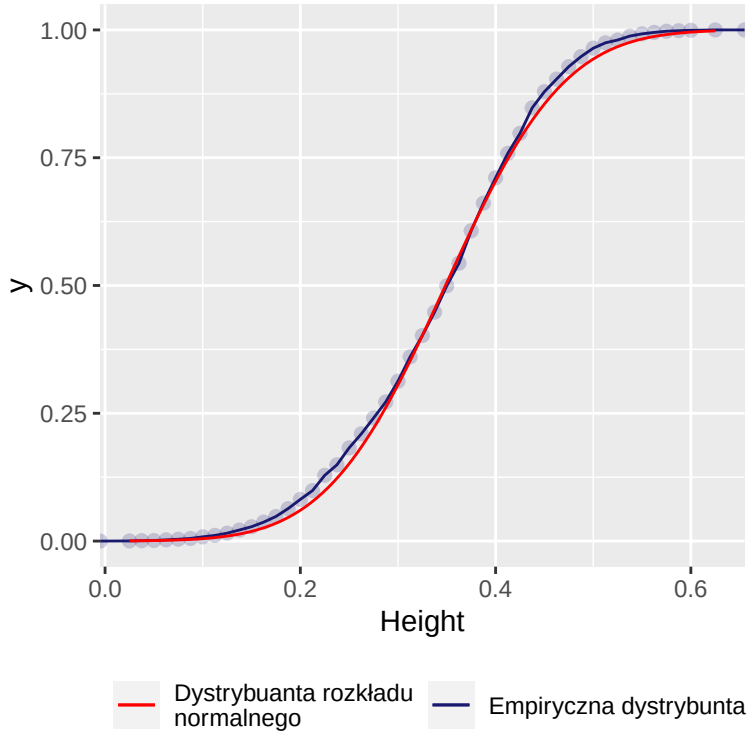
Rysunek 2.1. Wykres pudełkowy przed usunięciem wartości odstających



Rysunek 2.2. Wykres pudełkowy po usunięciu wartości odstających



Rysunek 2.3. Histogram wysokości



Rysunek 2.4. Dystrybuanta zmiennej objaśnianej

	Height
X	Min. :0.0250
X.1	1st Qu.:0.2875
X.2	Median :0.3625
X.3	Mean :0.3487
X.4	3rd Qu.:0.4125
X.5	Max. :0.6250
X.6	Variance: 0.0092
X.7	Skewness: -0.2707
X.8	Kurtosis: 2.8229
X.9	Variability: 0.2746

Tabela 2.2. Statystyki opisowe zmiennej objaśnianej po usunięciu obserwacji odstających

Patrząc na histogram 2.3 możemy stwierdzić, że dane są symetryczne i jednomodalne, moda wypada w okolicy wysokości równej 0.4 stopy. Co znaczy, że właśnie dla wartości zbliżonych do 0.4 wystpuje najwięcej obserwacji. Są to najbardziej popularne wysokości dla krabów z okolicy Bostonu. Na symetryczność rozkładu wskazują również średnia i mediana, które możemy odczytać z tabeli 2.2, obie mają bowiem zbliżone wartości. Mediana jest jedak trochę wyższa od średniej, co wskazuje na niewielką lewostronną skośność danych.

Na koniec policzymy jeszcze skośność (-0.2706531), kurtozę (2.8228722) i zmienność (0.2745705). Ujemna skośność, wskazuje na delikatną lewostronną asymetrię. Nie jest ona jednak duża. Kurtoza jest prawie równa 3, co świadczy o tym, że nasze dane mają rozkład zbliżony do normalnego. Na wykresie 2.3 przedstawiającym histogram i gęstość rozkładu normalnego oraz na wykesie 2.4 porównującym dystrybuanty, również widać, że różnica między rozkładem wysokości a rozkładem normalnym nie jest duża.

Rozrzut danych także jest niewielki, na co oprócz wariancji oraz rozstępu międzykwartylowego wskazuje przeciętna wartość współczynnika zmienności. Rozstęp międzykwartylowy możemy policzyć na podstawie danych z tabeli 2.2.

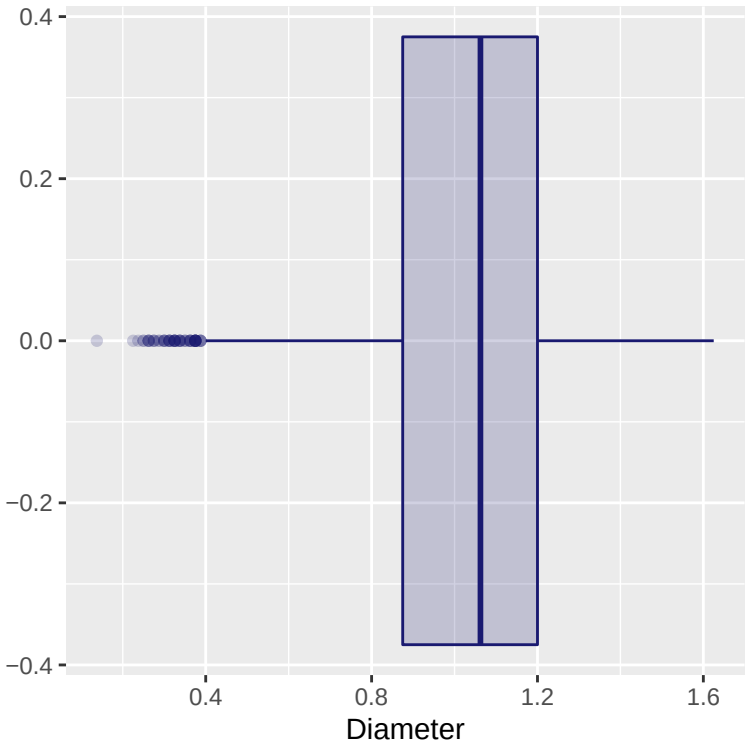
3. Analiza zmiennej objaśniającej

Przeanalizujemy zmienną objaśniającą, już po usunięciu obserwacji odstających. Podstawowe statystyki opisowe dla badanej zmiennej są wymienione w tabeli 3.1. Patrząc na tabelę 3.1 oraz na wykres pudełkowy 3.1. Widzimy, że wśród średnic krabów nie ma wartości mocno odstających bądź nietypowych, które trzeba by było usunąć.

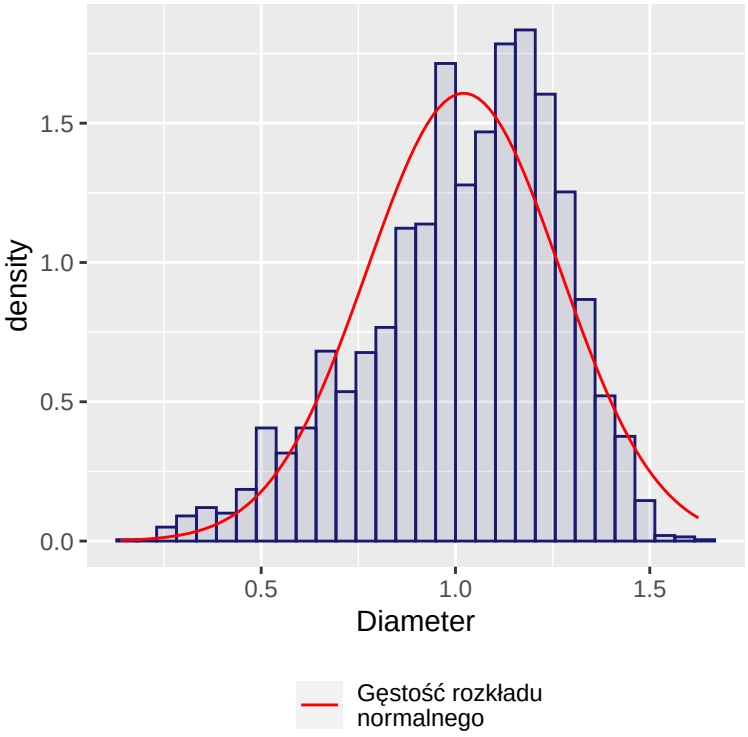
	Diameter
X	Min. :0.1375
X.1	1st Qu.:0.8750
X.2	Median :1.0625
X.3	Mean :1.0210
X.4	3rd Qu.:1.2000
X.5	Max. :1.6250
X.6	Variance: 0.0616
X.7	Skewness: -0.6189
X.8	Kurtosis: 2.9612
X.9	Variability: 0.2431

Tabela 3.1. Statystyki opisowe zmiennej objaśniającej

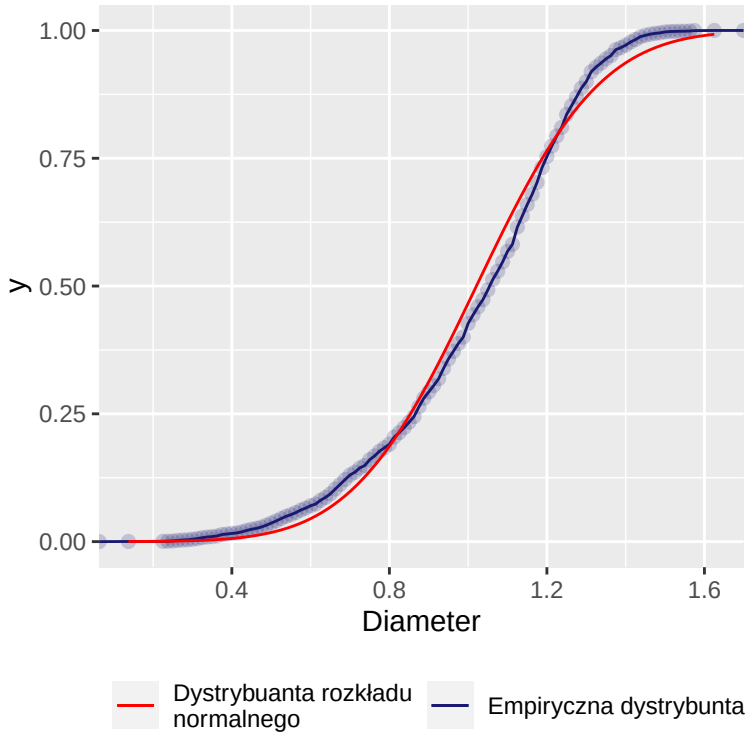
Ujemny współczynnik skośności świadczy o lewostronnej asymetrii. Jednak mimo tego zauważamy na wykresie dystrybuanty 3.3 oraz gęstości 3.2 podobieństwo do rozkładu normalnego. Kurtoza jest bliska 3, co również oznacza, że rozkład średnicy nie jest platykurtyczny, ani leptokurtyczny, lecz zbliżony do rozkładu normalnego. Podobnie jak dla zmiennej objaśnianej, mediana jest trochę wyższa od średniej, co ponownie wskazuje na lewostronną skośność. Różnica między medianą a średnią, nie jest jednak duża. Dla średnicy wariancja jest większa niż dla wysokości, co oznacza, że wartości leżą dalej od średniej. Wszystkie wymienione statystyki zmiennej objaśnianej można odczytać z tabeli 3.1.



Rysunek 3.1. Wykres pudełkowy zmiennej objaśniającej



Rysunek 3.2. Histogram średnicy



Rysunek 3.3. Dystrybuanta zmiennej objaśniającej

4. Regresja liniowa

Na wykresie 4.1 widać silną zależność liniową pomiędzy wysokością a średnicą.

W celu określenia zależności policzmy kowariancję oraz współczynnik korelacji Pearsona.

$$cov(X,Y) = E(XY) - EXEY$$

$$corr(X,Y) = \frac{cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Kowariancja wynosi 0.0215737, natomiast korelacja 0.9080893

Zatem występuje dodatnia zależność liniowa między zmienną objaśniającą a objaśnianą.

Spróbujemy dopasować do nich model regresji liniowej. Podzielimy dane na dwie części na 80% z nich zbudujemy model, by następnie przewidzieć pozostałe obserwacje.

Regresję liniową możemy zapisać jako:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

gdzie y_i - zmienna objaśniana, x_i - zmienna objaśniająca, ϵ_i - szum. Estymatory współczynników prostej wyznaczone metodą najmniejszych kwadratów mają postać:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})\ (x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

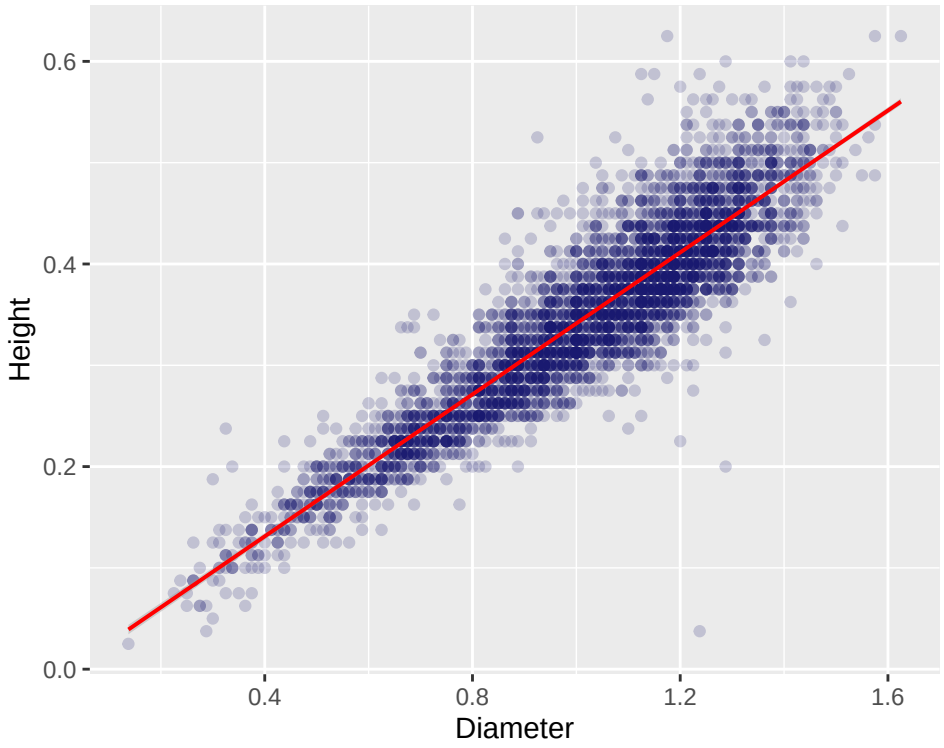
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\ \bar{x},$$

gdzie $\bar{y}$ - średnia zmiennej objaśnianej, $\bar{x}$ - średnia zmiennej objaśniającej. W naszym przypadku $\hat{\beta}_1 = 0.3504673$, $\hat{\beta}_0 = -0.0089128$.

Przedziały ufności na poziomie istotności α i biorąc pod uwagę, że σ jest nieznana mają postać:

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

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



Rysunek 4.1. Wykres rozproszenia wysokości od średnicy

gdzie $t_{1-\alpha/2,n-2}$ - kwantyl rzędu $1-\frac{\alpha}{2}$ rozkładu T-Studenta z n-2 stopniami swobody (n - długość próby, $\alpha = 0.05$), s - estymator odchylenia standardowego wyliczony ze wzoru:

$$s = \sqrt{\frac{1}{n-1}\sum(y_i - \hat{y}_i)^2}.$$

W naszym przypadku dla $\hat{\beta}_1$ mamy $[0.3448494, 0.3560851]$,
a dla $\hat{\beta}_0$: $[-0.0148098, -0.0030158]$.

Obliczmy również SST (Total Sum of Squares), SSE (Error Sum of Squares) oraz SSR (Regression Sum of Squares).

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

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

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

Dla naszych danych $SST = 28.4798389$, $SSE = 4.9036433$, $SSR = 23.5761955$.

Zatem współczynnik determinacji R^2 wynosi:
$\frac{SSR}{SST} = 0.8278205$.

Mówi on o jakości dopasowania, a dokładniej o tym jak blisko jest pula punktów od prostej regresji. Im większy ten paramter tym lepsze dopasowanie ($0 \leqslant R^2 \leqslant 1$). W naszym przypadku dopasowanie jest dość dobre.

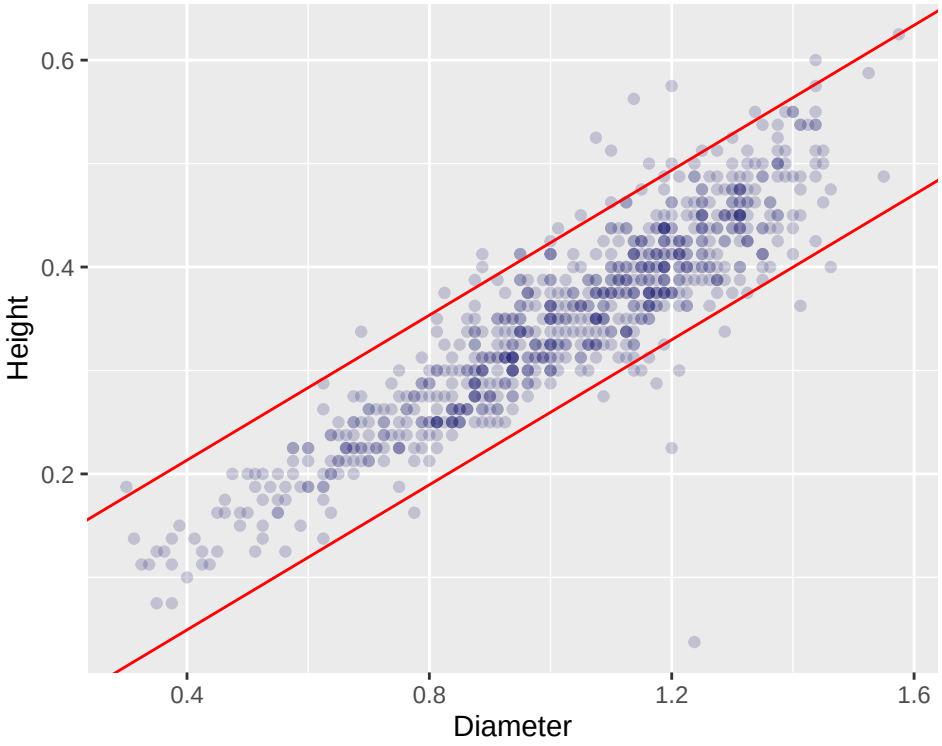
Na podstawie naszego modelu, spróbujemy teraz przewidzieć pozostałe 20% danych. Przedziały ufności dla nowej części danych możemy zapisać jako:

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

gdzie $t_{1-\alpha/2,n-2}$ - ponownie kwantyl rzędu $1-\frac{\alpha}{2}$ rozkładu T-Studenta z n-2 stopniami swobody (n - liczba obserwacji w zbiorze testowym, $\alpha = 0.05$), $\bar{x}$ - średnia zmiennej objaśniającej, $\hat{y}(x_0)$ - wyestymowana wysokość dla danej średnicy, s - estymator odchylenia standardowego wyliczony ze wzoru:

$$s = \sqrt{\frac{1}{n-2}\sum(y_i - \hat{y}_i)^2}.$$

Przedział ufności ma taką postać, dla każdego x_0 ze zbioru testowego. Wykres rozproszenia wysokości od śrenicy wraz z zaznaczonymi przedziałami ufności znajduje się na wykresie 4.2. Jak widać większość punktów mieści się w przedziale ufności, co znaczy, że nasz model regresji liniowej dobrze przewiduje nowe obserwacje.



Rysunek 4.2. Wykres rozproszenia wysokości od średnicy dla danych testowych i przedziały ufności

5. Analiza residuów

Założenia o residuach $\{\epsilon_i\}_{i=1,2,...,n}$:

— $\forall_{i=1,...,n}\ \epsilon_i \backsim N(0, \sigma^2)$,
— $E\epsilon = 0$,
— $\epsilon_1, ..., \epsilon_n \rightarrow$ niezależne,
— $Var\epsilon_i = \sigma^2 \quad \forall_{i=1,...,n}$ - wariancja jest stała.

5.1. Normalność

W celu sprawdczenia czy resiuda mają rozkład normalny przyjrzyjmy się najpierw wykresom. Po spojrzeniu na histogram 5.1 i wykres dystrybuanty 5.2 nie widać dużej rozbieżność z rozkładem normalnym. Jednak wykres kwantylowy 5.3 wskazuje na brak normalnośći rozkładu residuów, widoczne sa różnice w ogonach.

Następnie policzmy kurtozę - 5.012309. Dla rozkładu normalnego powinna ona być równa 3. W naszym przypadku jest jednak większa, zatem residua mają rozkład leptokurtyczny. Oznacza to, że intensywność wartości skrajnych jest większa niż dla rozkładu normalnego (ogony są cięższe).

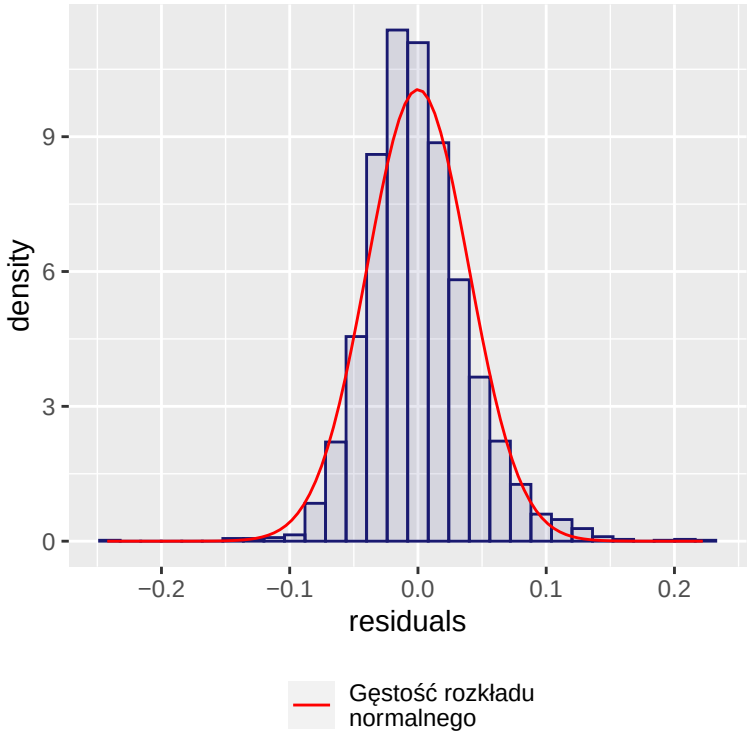
Powyższe metody nie potwierdzaja w pełni zgodności rozkładu residuów z rozkładem normalnym. W celu dalszej analizy wykonamy test Kołomogorwa-Smirnowa, Andersona-Darlinga oraz Jarque-Bera.

	test	p.value
1	Kołogorwa-Smirnowa	0.00
2	Andersona-Darlinga	0.00
3	Jarque-Bera	0.00

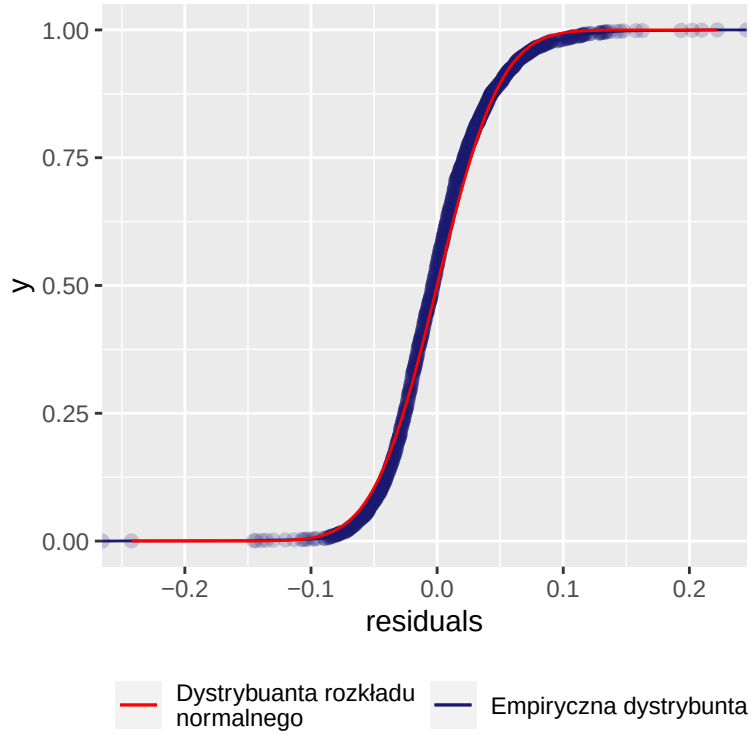
Tabela 5.1. Testy sprawdzające normalność residuów

Z tabeli 5.1 możemy odczytać, że dla każdego z powyższych testów p wartość jest bardzo bliska 0 (są one tak małe, że zostały zaokrąglone przez pakiet statystyczny do zera). Oznacza to, że odrzucamy hipotezę zerową o normalności residuów na rzecz hipotezy alternatywnej.

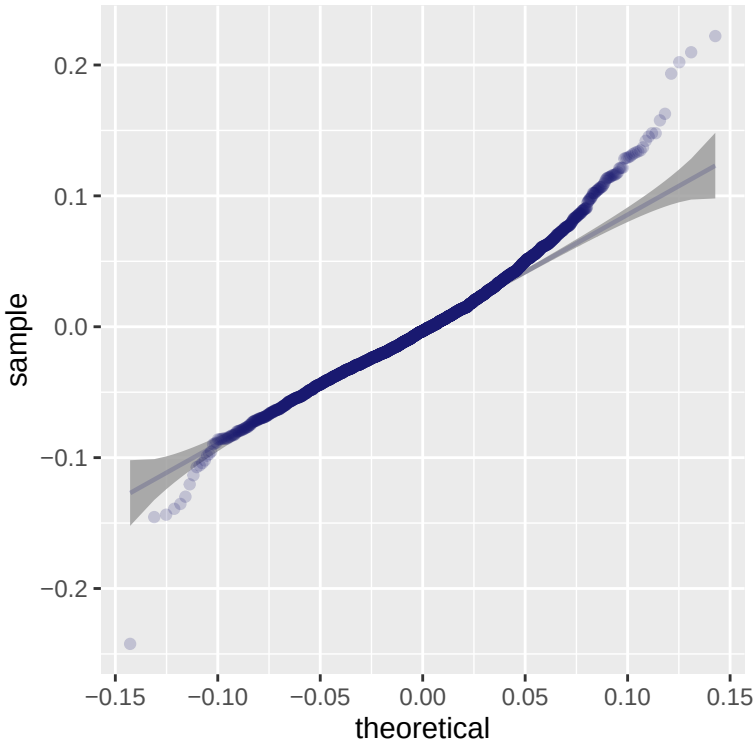
Możemy założyć, że reszty nie mają rozkładu normalnego. Nie jest, więc spełnione jedno z założeń regresji liniowej. Mimo wszystko przeprowadzimy analizę pozostałych założeń.



Rysunek 5.1. Histogram residuów oraz gęstość rozkładu normalnego



Rysunek 5.2. Dystrybuanta rozkładu normalnego oraz empiryczna residuów



Rysunek 5.3. Wykres kwanytlowy residuów

5.2. Średnia

Aby sprawdzić czy $E\epsilon = 0$ narysujmy najpierw wykres rozproszenia.

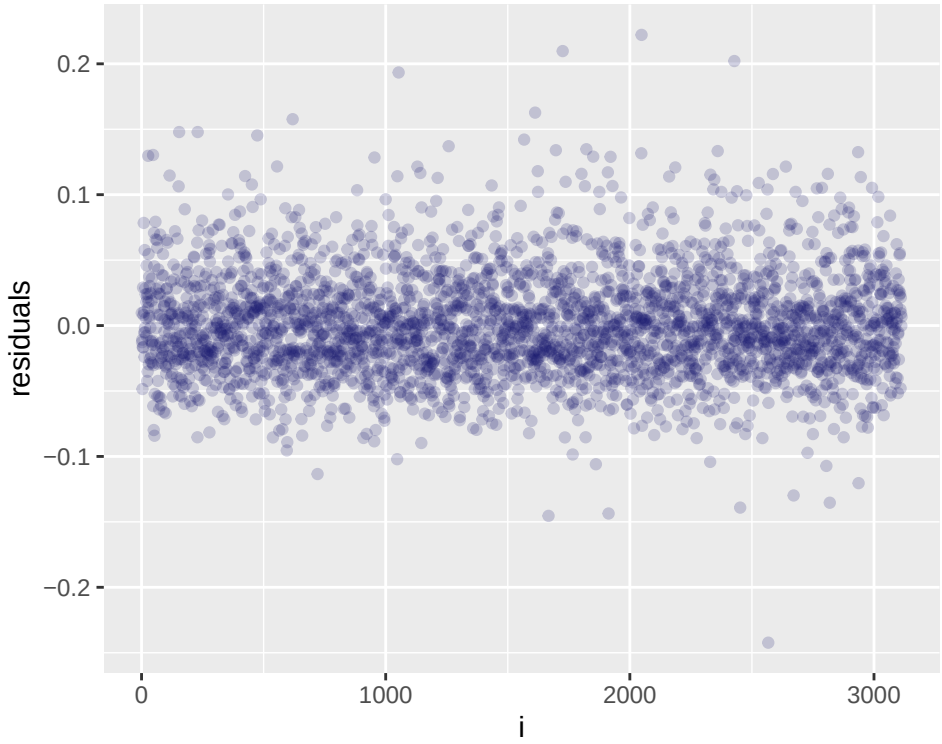
Z wykresu 5.4 widzimy że ϵ_i oscyluja blisko 0. Średnia próbkowa wynosi $-1.2186216 \times 10^{-18}$, więc jest bardzo blizka zeru. Wykonajmy jednak jeszcze test dla średniej (t.test). P wartość dla t testu, wynosi: 1. Nie mamy podstaw do odrzucenia hipotezy zerowej, więc możemy założyć, że średnia jest równa 0.

5.3. Niezależność reszt

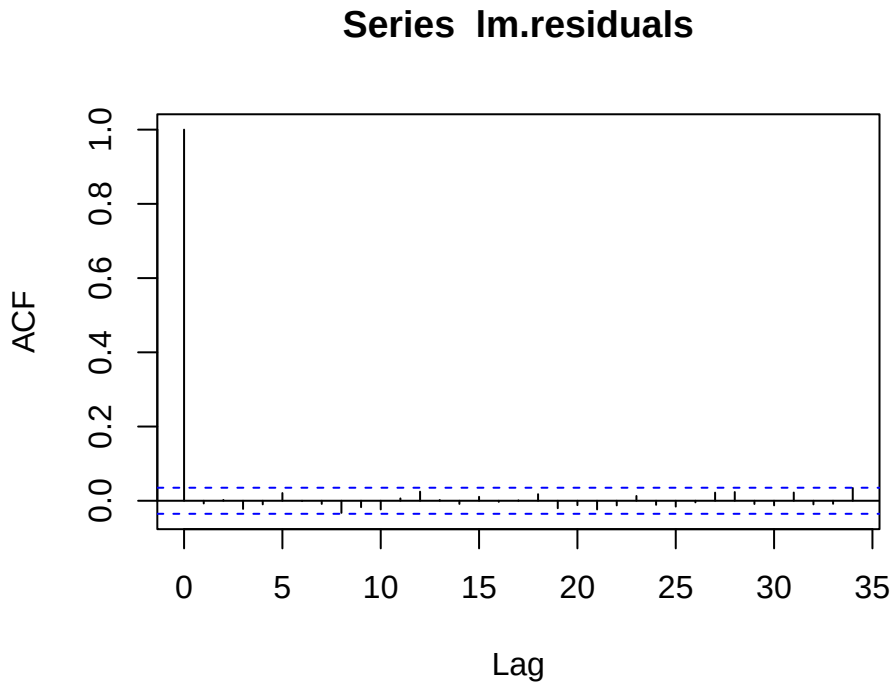
Sprawdzimy teraz, czy residua są niezależne. Zaczniemy od narysowania wykresu autokorelacji. Jak widać na wykresie 5.5, wszystkie punkty oscylują wokół zera i mieszczą się w przedziale ufności. Oznacza to, że residua nie są względem siebie skorelowane.

Następnie przeprowadzimy test Durbina-Watsona, który róznież pozwala ocenić czy występuje autokorelacja wśród reszt. P wartość w naszym przypadku wynosi 0.6569096. Oznacza to, że przy poziomie istotności 0.05 nie ma podstaw do odrzucenia hipotezy zerowej o braku korelacji między residuami. Możemy, więc założyć, że są one nieskorelowane. Podobny wynik uzyskujemy dla testu Ljunga-Boxa, gdzie p wartość jest równa 0.6856345.

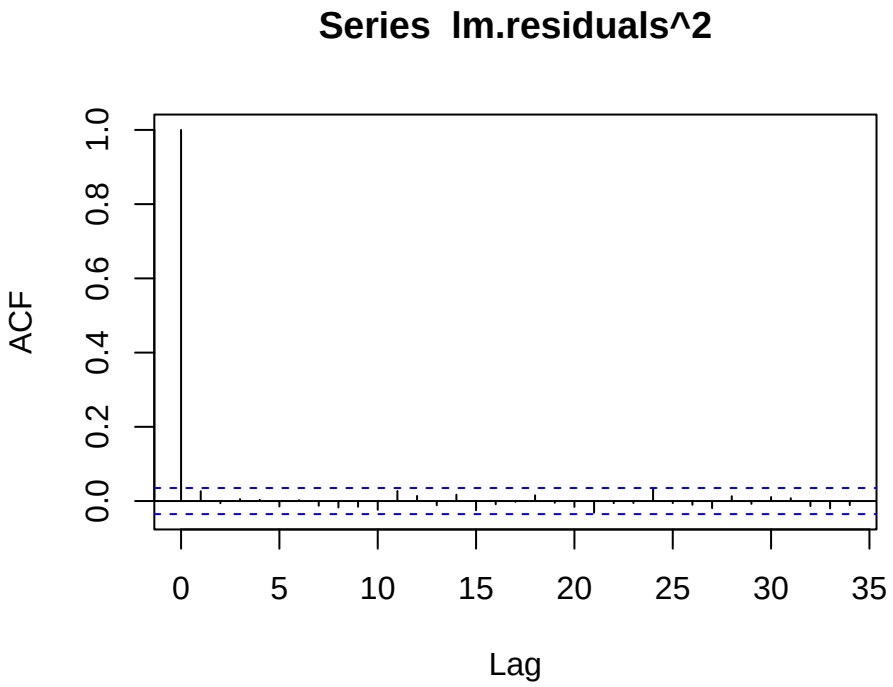
Wszystko, więc wskazuje na to, że założenie o braku korelacji między resztami jest spełnione, co znaczy, że ich rozkład jest losowy bez stale występującego wzorca.



Rysunek 5.4. Wykres rozproszenia residuów



Rysunek 5.5. Funkcja autokorelacji residuów



Rysunek 5.6. Funkcja autokorelacji residuów podniesionych do kwadratu

5.4. Stałość wariancji

Na koniec sprawdzimy założenie o stałości wariancji. Wykres 5.6 przedstawia autokorelację reszt podniesionych, do kwadratu. Jeśli residua podniesione do kwadratu są nieskorelowane, oznacza to, że są one homoskedastyczne, czyli mają stałą wariancję. Potwierdza to arch test, dla którego p wartość wynosi: 0.6282844. Jest ona większa od ustalonego poziomu istotności $\alpha = 0.05$. Możemy więc założyć, że hipoteza zerowa jest prawdziwa i residua są homoskedastyczne.

6. Podsumowanie

Zauważając silną liniową zależność między wysokością, a średnicą krabów z okolicy Bostonu, zdecydowałyśmy się wykonać model regresji liniowej na 80% danych. Do oszacowania parametrów modelu wykorzystałyśmy metodę najmniejszych kwadratów. Wyliczone na tej podstawie współczynnikiprostej regresji wynoszą: $\hat{\beta}_1 = 0.3504673$ i $\hat{\beta}_0 = -0.0089128$. Nasz model w dobry sposób predykował pozostałe dane. Zdecydowana większość wartości mieściła się w przedziałach ufności. Następnie sprawdziłyśmy założenia modelu. Niestety jeden z warunków - o normalności rozkładu residuów, nie został spełniony. W takim przypadku należy ponownie oszacować parametry modelu, stosując inną metodę estymacji albo inną postać modelu. Przykładowo, można zbudować model regresji liniowej dla danych po transformacji zmiennej zależnej przez logarytm lub pierwiastek.