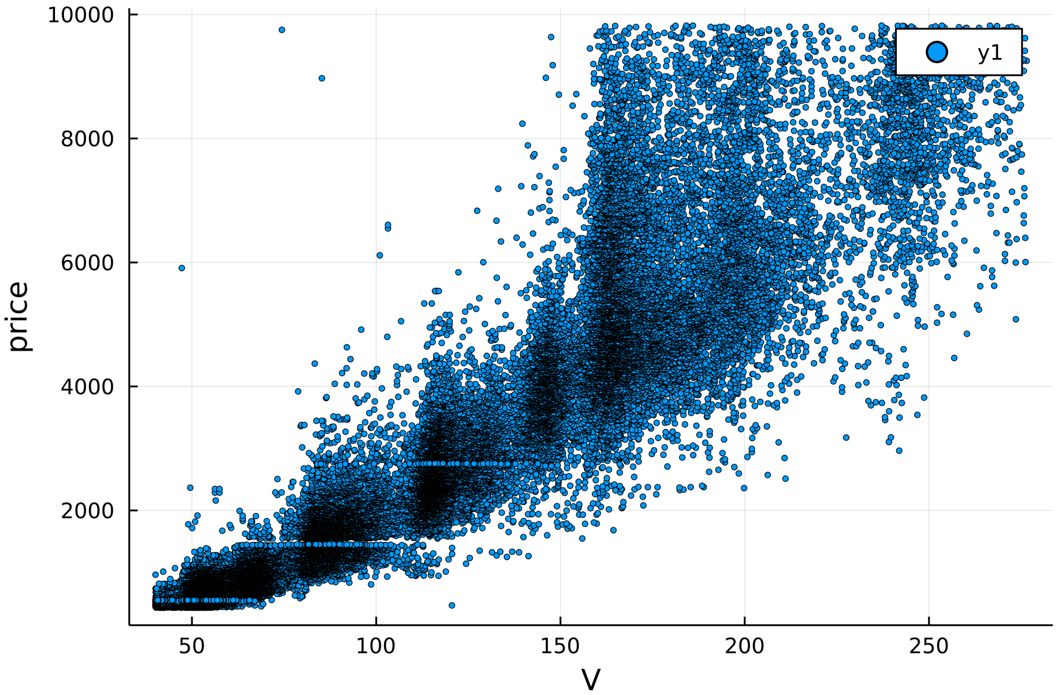
1 Analiza zależności między

Analizujemy zależność między ceną, a objętością. Przyjrzyjmy się zaleności naszych danych na wykresie.



Rysunek 1: Wykres cena-objętość

Z wykresu możemy zauważyć silną zależność między naszymi danymi. Z powodu rozłożenia naszych danych zależność ta może być linowa. W celu określenia tej zależności obliczymy współczynnik korelacji pearsona

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \approx 0.93,$$

gdzie x_i, y_i to i-te obserwacje odpowiednio objętości oraz ceny, $\overline{x}$ oznacza średnią z x, a n to rozmiar próby, oznaczeniami tymi będziemy posługiwali się w całym raporcie. Wartość ta jest blisko wartości 1, zatem nasze dane są silnie skorelowane dodatnie oraz liniowo. Dlatego model, którym będziemy opisywać dane będzie miał postać

$$Y_i = \beta_0 + \beta_1 x_i + \xi_i, \tag{1}$$

gdzie ξ_i jest losowym błędem pomiarowym. Y_i jest zmienną losową, której realizacją jest zaobserwowana wartość y_i. Zgodnie z treścią polecenia zakładamy, że wszystkie zminne losowe ξ_i są idd. oraz mają rozkład normalny ze średnią 0. Naszym zadaniem będzie estymować wartość $\hat{y}_i$ poprzez estymowanie realizacji zmiennych losowych $\hat{\beta}_0$ oraz $\hat{\beta}_1$.

1.1 Estymacja punktowa

By stworzyć estymator ceny $\hat{y}$ skorzystamy z metody najmniejszych kwadratów. Do estymacji parametrów β_0 oraz β_1, występujące we wzorze (1), wykorzystamy losowo wybrane 80% naszych danych. Pozostałe

20% posłuży nam w celu sprawdzenie poprawności modelu. Estymowane parametry, w danej realizacji Y, mają wartość

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \approx 39.255 \quad \text{oraz} \quad \hat{\beta}_0 = \overline{y} - \beta_1 \overline{x} \approx -1564.73.$$

Estymowana wartośc ceny $\hat{y}$ w modelu (1) będzie miała postać

$$\hat{y}_i = 39.255 \cdot x_i - 1564.73.$$

1.2 Estymacja przedziałowa

W tym przypadku będziemy szukać przedziału do w którym będzie należał nasz Y_i z dużym prawdopodobieństwem. Będziemy chcieli by prowdopodobieństow to wynosiło $1 - \alpha$. Znany jest fakt, że w modelu (1) zmienna $\hat{Y}_i$ ma poniższe parametry

$$\mathbb{E}\hat{Y}_i = \beta_0 + \beta_1 x_i \quad \text{oraz} \quad Var\hat{Y}_i = Var(\xi_1)\left(\frac{1}{n} + \frac{(x_0 - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right).$$

Jeśli $Var(\xi_1)$ nie jest znana w jej miejsce wstawiamy jej estymator

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)$$

otrzymując estymator $Var\hat{Y}_i$ i wtym przypadku zmienna losowa

$$\frac{\hat{Y}_i - \mathbb{E}\hat{Y}_i}{s^2\left(\frac{1}{n} + \frac{(x_0 - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}$$

ma rozkład t-studenta z $n - 2$ stopniami swobody. Niech $z_{\alpha/2}$ będzie $1 - \alpha/2$ kwantylem z tego rozkładu. Wtedy możemy napisać, że

$$\mathbb{P}\left(\hat{Y}_0 - z_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \leqslant Y_0 \leqslant \hat{Y}_0 + z_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\right) = 1 - \alpha$$