Regresja liniowa

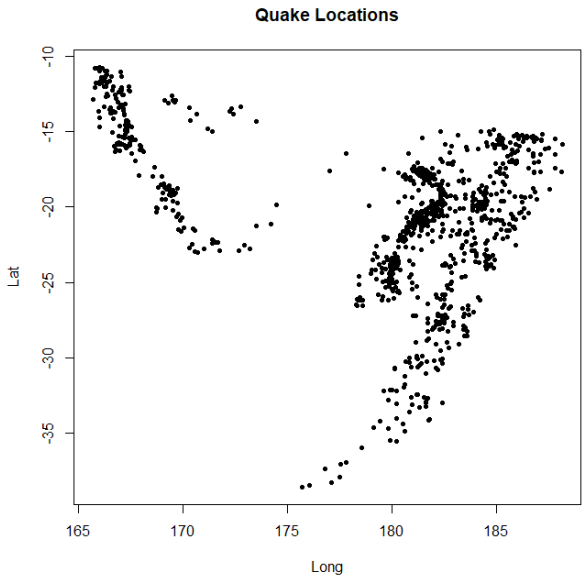
Simple Linear Regression (SLR)

W prostym modelu regresji liniowej interesuje nas, czy istnieje liniowa zależność pomiędzy zmienną objaśniającą (czasami nazywaną zmienną niezależną) a zmienną objaśnianą (zależną). W tym ćwiczeniu będziemy sprawdzać, czy istnieje liniowa zależność pomiędzy wielkością trzęsienia ziemi a liczbą stacji, które zgłosiły aktywność.

Model regresji liniowej

$$y = a_0 + a_1 x + \varepsilon$$

1. Zaimportuj pakiet **csrplus**. Wczytaj zbiór danych **quakes**, który zawiera informacje o lokalizacji trzęsień ziemi zarejestrowanych w rejonie Fidżi od 1964 roku o magnitudzie >4.0. Zbiór ten składa się z 1000 obserwacji i jest opisany przez 5 zmiennych- długość geograficzną, szerokość geograficzną, głębokość, magnitudę zjawiska oraz ilość stacji które zarejestrowały dane trzęsienie. Przedstaw mapę występowania epicentrów w badanym obszarze.



2. Przedstaw na wykresie punktowym (scatter plot) zależność pomiędzy ilością stacji które zarejestrowały zjawisko a magnitudą tego trzęsienia.

   Jedną z rzeczy, którą można zauważyć po wyświetleniu danych jest prawie kolumnowy układ. Wynika to najprawdopodobniej z faktu, że skala magnitudy nie jest ciągła, a raczej zapisywana w odstępach co 0,1. Można również zauważyć, że gdy istnieje wiele podobnych wartości, nakładają się one na siebie, co utrudnia nam zrozumienie prawdziwego rozkładu.

3. Aby lepiej zrozumieć dane, dodaj szum losowy do naszych wartości magnitudy za pomocą funkcji **jitter()** podając maksymalną wartość o jaką mogą się przemieścić punkty= 0.5. Przy wyświetlaniu wykresu, ustaw pch=20 oraz kolor RGB= [0.1, 0.2, 0.8] o parametrze

przezroczystości alfa=0.3, co oznacza że punkt będzie zacieniony na ciemniejszy kolor, gdy gęstość w danym punkcie wzrasta.

Wystarczy spojrzeć na wykres rozproszenia, aby zauważyć, że wraz ze wzrostem magnitudy trzęsienia ziemi wzrasta również liczba stacji, które je rejestrują. Możemy również zwrócić uwagę na rozprzestrzenianie się danych lub na to, jak blisko siebie znajdują się punkty danych.

4. Utwórz model regresji liniowej o nazwie **Quake.mod** za pomocą funkcji **lm()**. Jako argumentów funkcji należy użyć zmiennej objaśnianej i objaśniającej.

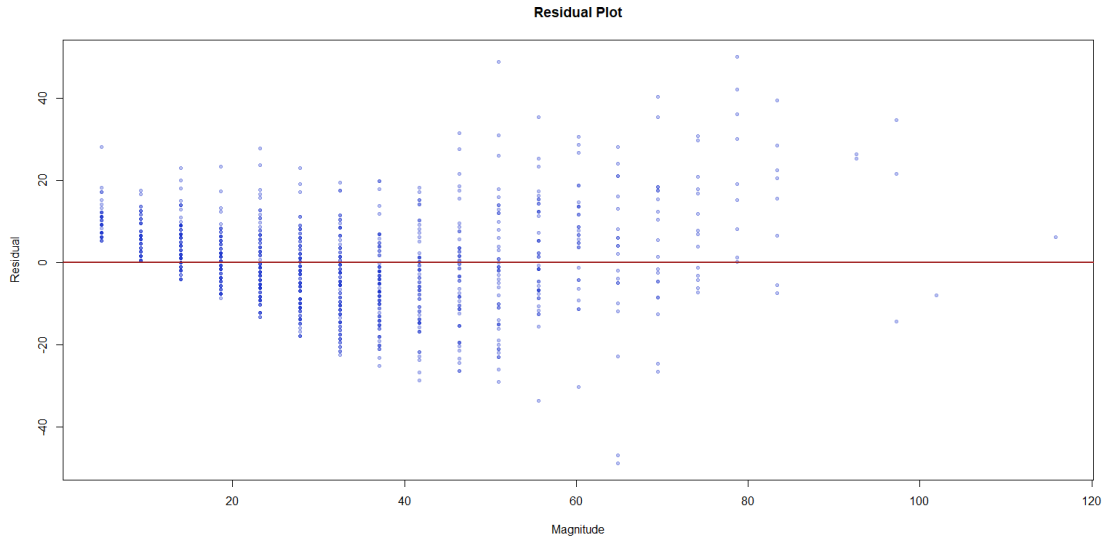
   Model przedstawia liczbową zależność, opartą na naszych przykładowych danych, pomiędzy wielkością trzęsienia ziemi a liczbą stacji, które zarejestrowały trzęsienie. Ze współczynnika nachylenia (slope coefficient) dowiedzieliśmy się, że zmiana o 1 w skali Richtera spowoduje zmianę liczby stacji zgłaszających średnio o 46,28. Ponieważ nasze nachylenie jest dodatnie, nasz model przewiduje, że istnieje pozytywny związek pomiędzy wielkością i liczbą stacji, które zgłaszają trzęsienie ziemi. Intercept mówi nam, że gdyby wielkość trzęsienia ziemi była równa zeru, - 180,42 stacji zgłosiłyby to trzęsienie. To oczywiście nie ma sensu, więc jeżeli zbiór danych nie zawiera wartości dla zera lub około zera, nie można zinterpretować współczynnika intercept. Wartość ta może być również interpretowana tylko wtedy, gdy jest rozsądna i prawdopodobna.

5. Dodaj na wykresie punktowym linię regresji.

   Założenia analizy regresji liniowej

   Jednym z głównych założeń, które należy sprawdzić aby móc zastosować model SLR jest **homoscedastyczność** - wariancja reszt (residuals), składnika losowego jest taka sama dla wszystkich obserwacji. Innymi słowy, wariancja naszych reszt jest niezależna od naszej zmiennej prognostycznej. Istnieje kilka sposobów, na które możemy sprawdzić wariancję naszych reszt.

6. Podstawową opcją sprawdzenia wariancji reszt jest wykres rezydualny z wartościami dopasowanymi naszego modelu na osi X i Y. Stwórz zmienną dla naszych reszt (**QuakeResiduals**) i wartości dopasowanych (**QuakeFittedValues**), a następnie zaznacz na wykresie reszt poziomą linią wartość Y równą zero jako punkt odniesienia dla zmian naszych reszt.



Kolejnym założeniem jest aby **reszty miały rozkład zbliżony do normalnego.** Cechy rozkładu normalnego:

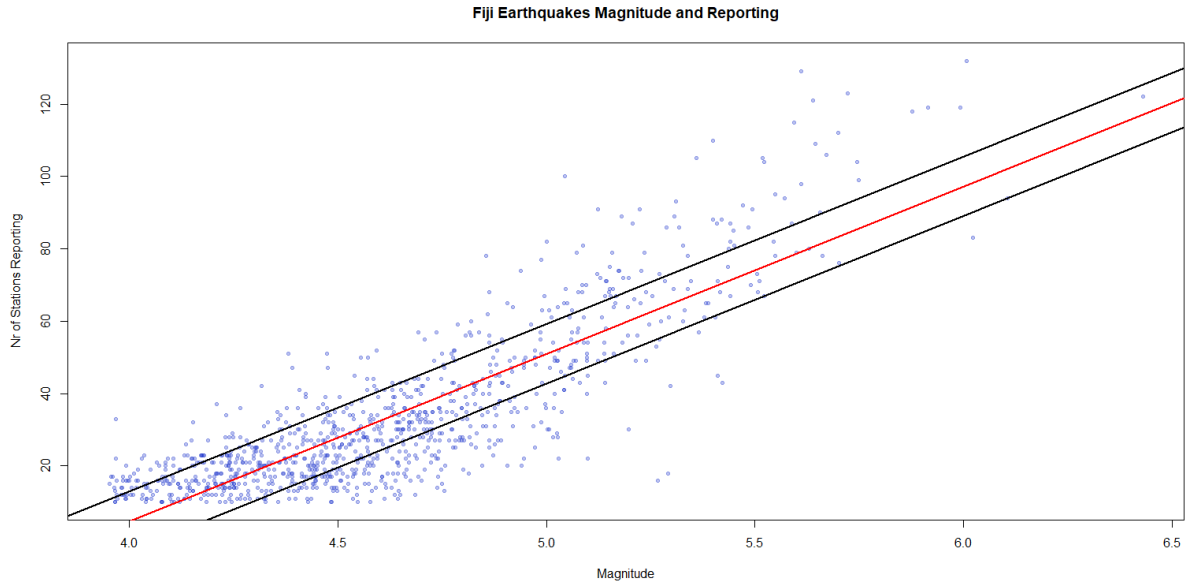
- Krzywa w kształcie dzwonu (bell-shaped curve)

- Większość danych w ramach jednego i dwóch odchyleń standardowych od średniej

- Symetryczny kształt

7. Przedstaw histogram reszt aby sprawdzić, czy mają rozkład normalny.

8. Użyj funkcji **confint()** do obliczenia przedziału ufności dla każdego ze współczynników regresji przyjmując współczynnik ufności równy 0.95. Przedstaw przedział ufności na wykresie.



9. Oceń **dokładność oszacowania modelu**. Oblicz:
   a) resztkową sumę kwadratów odchyleń (Residual Sum of Squares)

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

b) odchylenie standardowe składnika resztkowego (Residual Standard Error)

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

c) współczynnik zbieżności (Fraction of variance unexplained)

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

Wskazuje, jaka część zmienności zmiennej y nie jest objaśniona za pomocą oszacowanej funkcji regresji. Przyjmuje wartości z przedziału od 0 do 1. Im bliżej 0, tym lepsze dopasowanie funkcji regresji do danych empirycznych.

d) współczynnik determinacji (Coefficient of determination)

$$R^2 = 1 - \varphi^2$$

Określa stopień, w jakim oszacowana funkcja regresji wyjaśnia zmienność zmiennej y. Przyjmuje wartości z przedziału od 0 do 1. Im bliżej 1, tym lepsze dopasowanie funkcji regresji do danych empirycznych.

10. Wykonaj test istotności statystycznej dla współczynnika regresji liniowej. Celem testowania hipotez jest sprawdzenie, czy istnieje liniowa zależność między zmienną objaśniającą a zmienną objaśnianą. Testujemy to poprzez badanie nachylenia modelu regresji. Wyświetl funkcję **summary()** dla **Quake.mod** i na podstawie wartości p-value sprawdź, czy możemy odrzucić hipotezę zerową o nieistotności obu parametrów modelu regresji.