# Analiza wypadków w Polsce względem województw dla roku 2021

Żaneta Sado, Gabriela Ryszka

2024-01-07

## Spis Treści



# 1 Dane

Dane zostały pobrane ze strony https://stat.gov.pl, https://autostrady.info.pl/. Nasze zestawienie zawiera 16 obserwacji (względem województw) i 7 zmiennych:

-Wypadki->Wypadki ogółem

-Drogi->Długość dróg publicznych [km]

-Wykroczenia->Ilość kierowców zatrzymanych

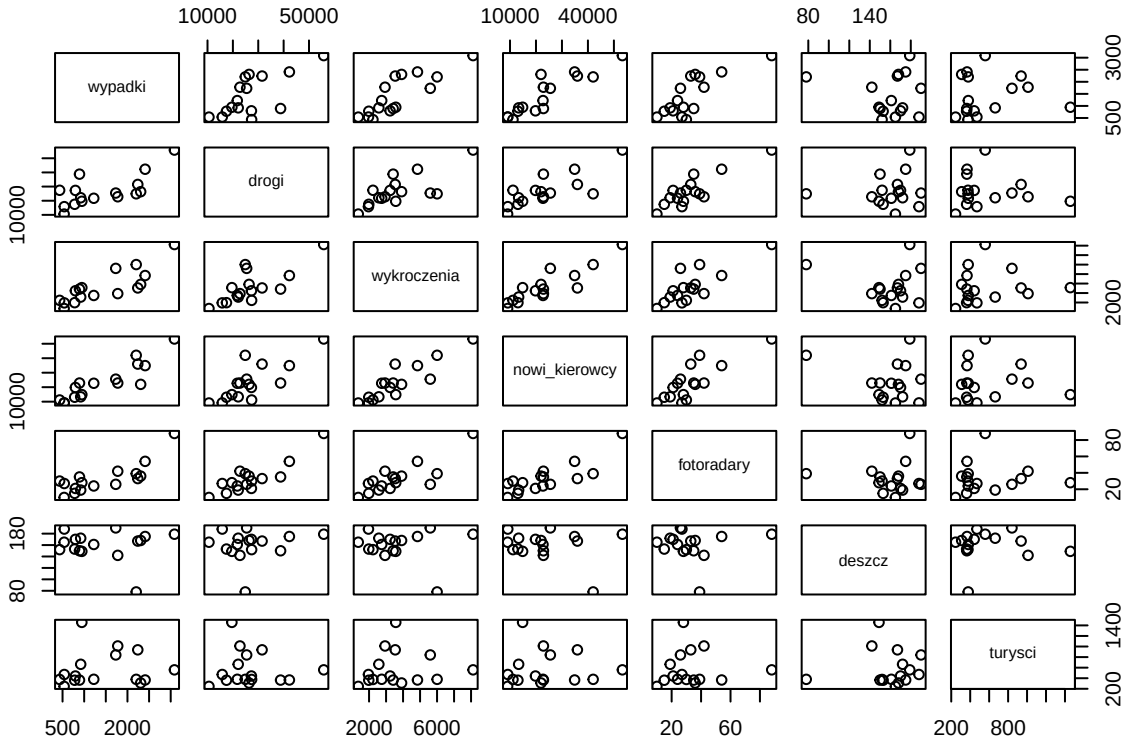
-Nowi_kierowcy->Ilość nowych kierowców

-Fotoradary->Ilość fotoradarów

-Deszcz->Średnia ilość dni opadów deszczu w ciągu roku

-Turyści->Ilość turystów na 1000 ludności

Wczytujemy teraz nasze dane:




Porównując do pozostałych, widzimy, że zmienne wykroczenia i turyści mają mniejsze zagęszczenie oraz sporo wartości odstających. Podobna sytuacja jest w przypadku turystów i nowych kierowców. Reszta zmiennych jest raczej zależna od siebie losowo, chociaż w przypadku fotoradarów poza zmiennymi deszcz i turyści zbliżona jest trochę do linii.

```
##     wypadki          drogi        wykroczenia   nowi_kierowcy
##  Min.   : 433.0   Min.   :10568   Min.   :1366   Min.   : 9076
##  1st Qu.: 792.8   1st Qu.:21203   1st Qu.:2489   1st Qu.:13262
##  Median :1084.0   Median :25152   Median :3328   Median :22324
##  Mean   :1426.0   Mean   :26863   Mean   :3628   Mean   :23278
##  3rd Qu.:2212.0   3rd Qu.:28420   3rd Qu.:4147   3rd Qu.:27873
##  Max.   :3086.0   Max.   :55810   Max.   :8119   Max.   :53272
##    fotoradary       deszcz         turysci
##  Min.   :10.00   Min.   : 78.0   Min.   : 245.5
##  1st Qu.:23.25   1st Qu.:151.5   1st Qu.: 366.2
##  Median :29.00   Median :166.0   Median : 411.0
##  Mean   :32.94   Mean   :159.9   Mean   : 572.9
##  3rd Qu.:36.75   3rd Qu.:172.8   3rd Qu.: 708.1
##  Max.   :88.00   Max.   :190.0   Max.   :1456.9
```

Przykładowo dla wykroczenia nasza minimalna wartość w kolumnie to 1366, a największa wartość to 8119. Mediana to 3328, a średnia 3628. Pierwszy kwantyl wynosi 2489, a trzeci kwantyl 4147.

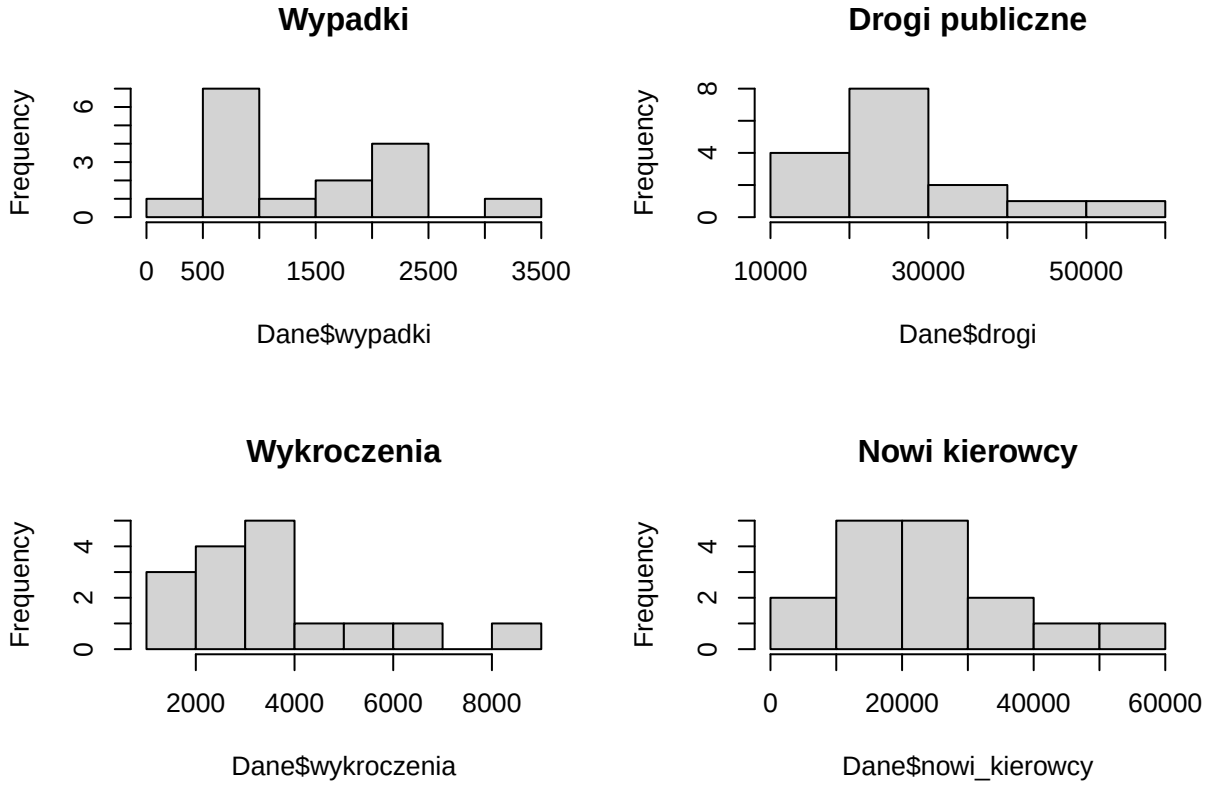
```
##       wypadki         drogi   wykroczenia nowi_kierowcy    fotoradary
##     0.4921381     1.1397354     1.1122215     0.9255051     1.7440278
##        deszcz       turysci
```

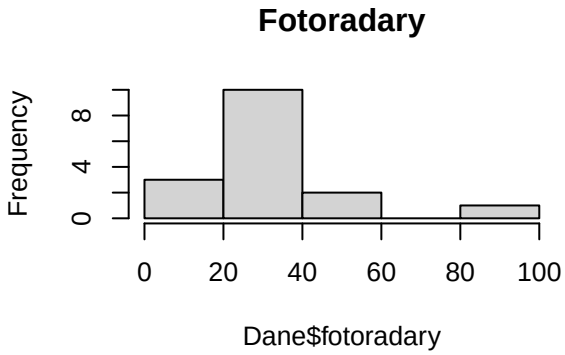
```
##    -1.9720938     1.4129678
```

W przypadku deszczu mamy skośność ujemną, co oznacza że ogon rozkładu jest wydłużony w lewo, co sugeruje że większość obserwacji w próbie ma wartości poniżej średniej. Dla pozostałych wartości mamy skośność dodatnią, co oznacza że ogon rozkładu jest wydłużony w prawo, co sugeruje że większość obserwacji w próbie ma wartości powyżej średniej.

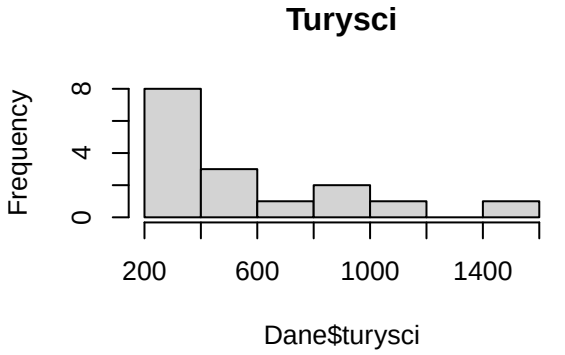
```
##       wypadki         drogi   wykroczenia nowi_kierowcy    fotoradary
##      1.966732      4.141494      3.739041      3.098183      6.329601
##        deszcz       turysci
##      7.446693      4.178665
```

W przypadku wszystkich zmiennych poza wypadkami kurtoza ma wartości większe od 3, zatem oznacza to leptokurtyczność czyli bardziej spiczasty rozkład w porównaniu do rozkładu normalnego. Sprójrzmy teraz na histogramy naszych zmiennych:





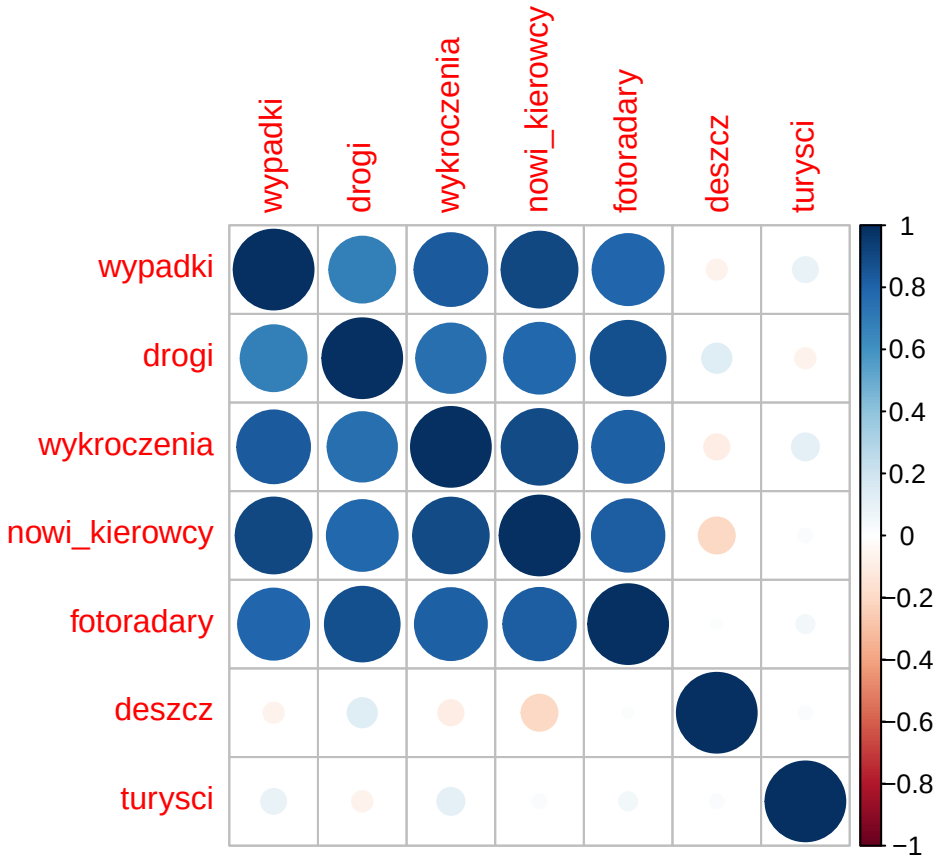




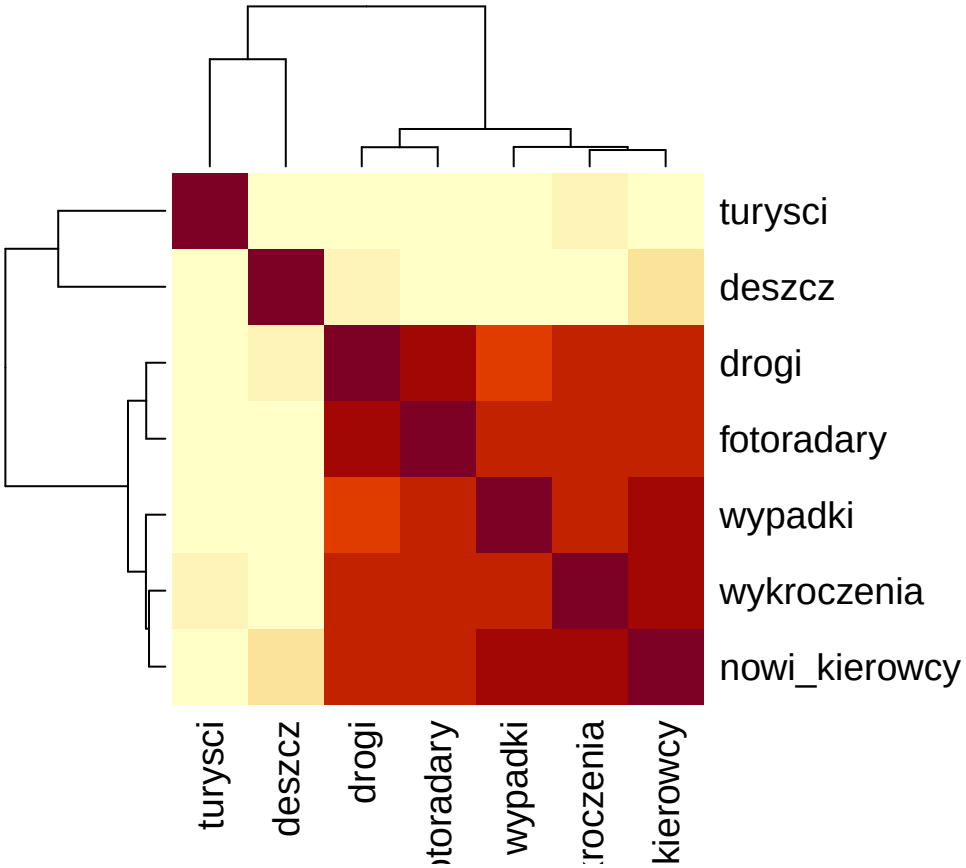

Widać że wcześniejsza analiza dla kurtozy i skośności potwierdza się na powyższych histogramach. Zobaczmy teraz, jak wygląda macierz korelacji pomiędzy zmiennymi egzogenicznymi.

```
##                    wypadki       drogi wykroczenia nowi_kierowcy fotoradary
## wypadki         1.00000000  0.68193270   0.8322295    0.90384707 0.79242952
## drogi           0.68193270  1.00000000   0.7554076    0.78700514 0.87522602
## wykroczenia     0.83222951  0.75540761   1.0000000    0.89796367 0.81217944
## nowi_kierowcy   0.90384707  0.78700514   0.8979637    1.00000000 0.82316298
## fotoradary      0.79242952  0.87522602   0.8121794    0.82316298 1.00000000
## deszcz         -0.06453889  0.13404138  -0.0993213   -0.20331483 0.01765759
## turysci         0.09761613 -0.06603328   0.1141853    0.02989685 0.05220817
##                     deszcz     turysci
## wypadki        -0.06453889  0.09761613
## drogi           0.13404138 -0.06603328
## wykroczenia    -0.09932130  0.11418529
## nowi_kierowcy  -0.20331483  0.02989685
## fotoradary      0.01765759  0.05220817
## deszcz          1.00000000  0.02844886
## turysci         0.02844886  1.00000000
```

Korelację ujemną dostaliśmy w przypadku opadów dla wszystkich zmiennych poza drogami i fotoradarami. Poza tym, ujemna jest również dla turystów i dróg. W pozostałych przypadkach wartości korelacji są dodatnie. Najsilniejszy współczynnik korelacji jest dla zmiennych: wykroczenia, nowi_kierowcy oraz fotoradary. Na podstawie tego zmienną zależną będą wypadki.




Widać tutaj, że najwięcej ciemnego niebieskiego jest dla zmiennej wykroczenia, nowi_kierowcy oraz fotoradary, zatem między tymi zmiennymi jest najsilniejsza korelacja. Najsłabsza korelacja jest zatem dla deszczu i turystów.




Tutaj ponownie potwierdzają się powyższe wnioski dla siły korelacji pomiędzy rozważanymi zmiennymi.

```
##
##  Pearson's product-moment correlation
##
## data:  Dane$wypadki and Dane$turysci
## t = 0.367, df = 14, p-value = 0.7191
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4183318  0.5659365
## sample estimates:
##        cor
## 0.09761613
```

Mamy tutaj do czynienia z korelacją między dwoma zmiennymi-wypadki i turyści. Statystyka wynosi 0.367, a p-value=0.7191 <0.05, co oznacza, że nie mamy podstaw do odrzucenia hipotezy zerowej.(H0: Korelacja między zmiennymi wynosi 0, H1: Rzeczywista korelacja między zmiennymi nie jest równa 0). Przedział ufności zawiera 0, więc dodatkowo potwierdza że nie ma znaczącej różnicy między zmienną wypadki a turyści.

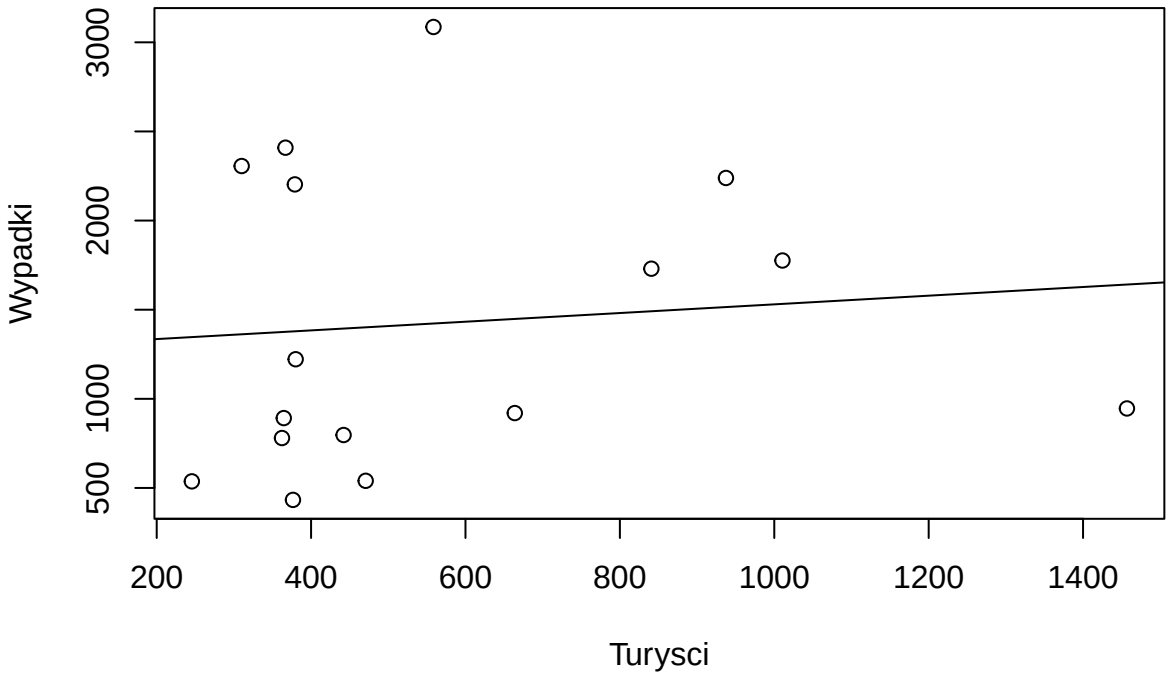
```
##
##  Pearson's product-moment correlation
##
## data:  Dane$wypadki and Dane$deszcz
## t = -0.24199, df = 14, p-value = 0.8123
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.5428764  0.4454163
## sample estimates:
##         cor
## -0.06453889
```

Tutaj natomiast zbadana została korelacja pomiędzy zmiennymi: wypadki, a deszcz. Test nie wykazuje tutaj silnego związku pomiędzymi tymi zmiennymi. Nie mamy podstaw do odrzucenia hipotezy zerowej (H0: Korelacja między zmiennymi wynosi 0, H1: Rzeczywista korelacja między zmiennymi nie jest równa 0).

## 2 Model pełny-regresja liniowa

Teraz dopasowany zostanie model pełny, będziemy na jego podstawie analizować czy możemy go jakoś przekształcić w celu jego ulepszenia.




Linia regresji nachyla się lekko w górę, zatem wraz ze wzrostem liczby turystów rośnie liczba wypadków. Jednak punkty wokół linii regresji są rozrzucone, co może sugerować że zmienna turysci nie ma znacznego wpływu na model.

```
##
## Call:
## lm(formula = wypadki ~ ., data = Dane)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -387.09 -155.00  -73.25   87.44  868.96
##
## Coefficients:
##                  Estimate  Std. Error t value Pr(>|t|)
## (Intercept)   -725.706687  738.611178  -0.983    0.352
## drogi           -0.031143    0.021354  -1.458    0.179
## wykroczenia     -0.001868    0.137439  -0.014    0.989
## nowi_kierowcy    0.065034    0.021685   2.999    0.015 *
## fotoradary      15.235822   13.101967   1.163    0.275
## deszcz           5.966348    4.544061   1.313    0.222
## turysci          0.043989    0.317429   0.139    0.893
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 384.4 on 9 degrees of freedom
## Multiple R-squared:  0.8691, Adjusted R-squared:  0.7819
## F-statistic: 9.961 on 6 and 9 DF,  p-value: 0.001514
```

Nasz pełny model to:

Wypadki=-725,71-0,03*drogi-0,00187*wykroczenia+0,065*nowi_kierowcy+15,24*fotoradary+5,97*deszcz+0,044*turysci.

Duża wartość p świadczy, że zmienne wykroczenia i turyści są nieistotne. Zmienność wypadków jest wyjaśniana przez nasz model w około 87%. $R^2$ poprawiony wynosi 78%. Spróbujmy zobaczyć co się stanie gdy usuniemy z naszego modelu zmienne wykroczenia i turysci.

```
##
## Call:
## lm(formula = wypadki ~ . - turysci - wykroczenia, data = Dane)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -397.07 -169.11  -57.49  103.27  855.04
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -707.66515  658.36284  -1.075    0.305
## drogi           -0.03191    0.01867  -1.709    0.116
## nowi_kierowcy    0.06504    0.01466   4.437    0.001 **
## fotoradary      15.53458   11.41466   1.361    0.201
## deszcz           6.03531    4.06449   1.485    0.166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 348.1 on 11 degrees of freedom
## Multiple R-squared:  0.8688, Adjusted R-squared:  0.8211
## F-statistic: 18.22 on 4 and 11 DF,  p-value: 0.00008124
```

Zmienność wypadków jest wyjaśniana przez nasz model w około 87%. Natomiast $R^2$ poprawiony wynosi 82%, co wskazuje znaczną poprawę.

# 3 Regresja wieloraka

Wracamy teraz do analizy pełnego modelu:

```
##                       2.5 %       97.5 %
## (Intercept)   -2396.56125344 945.14787902
## drogi            -0.07945020   0.01716339
## wykroczenia      -0.31277724   0.30904150
## nowi_kierowcy     0.01597884   0.11408835
## fotoradary      -14.40288528  44.87452956
## deszcz           -4.31303153  16.24572846
## turysci          -0.67408555   0.76206452
```

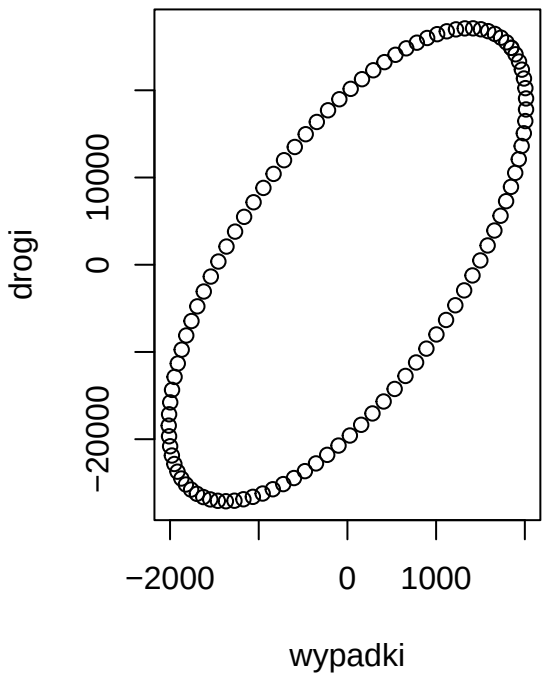
Przedział ufności dla wykroczenia od -0,313 do 0,31 jest to jeden z wielu przedziałów które z prawdopodobieństwem 95% zawierają nieznaną wartość parametru wykroczenia.

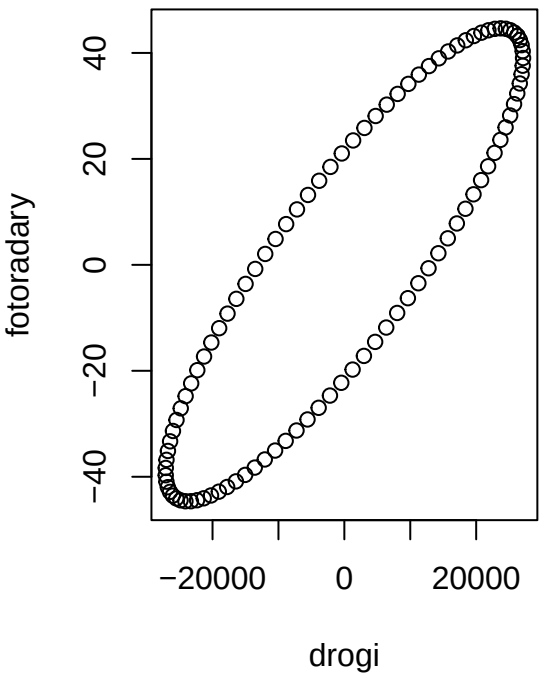
```
##                         5 %          95 %
## (Intercept)   -2079.66438932 628.251014898
## drogi            -0.07028826   0.008001447
## wykroczenia      -0.25380969   0.250073943
## nowi_kierowcy     0.02528264   0.104784549
## fotoradary       -8.78156214  39.253206419
## deszcz           -2.36342841  14.296125329
## turysci          -0.53789432   0.625873298
```

Przedział ufności dla wykroczenia od -0,253 do 0,25 jest to jeden z wielu przedziałów które z prawdopodobieństwem 90% zawierają nieznaną wartość parametru wykroczenia.

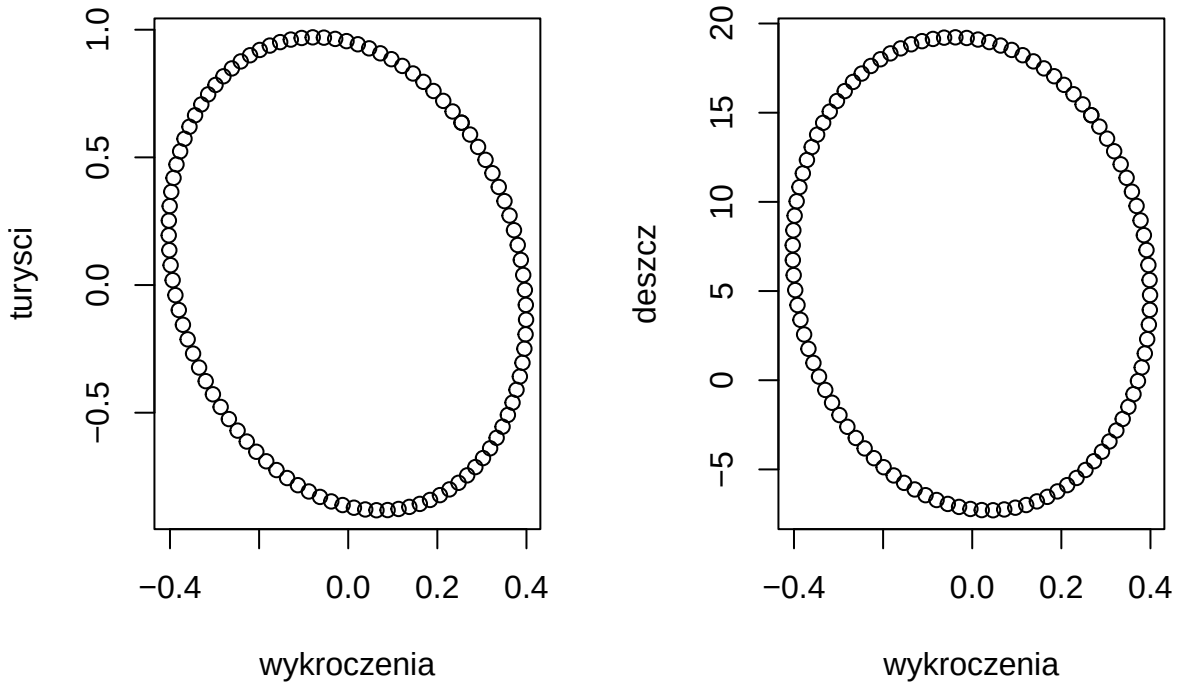
```
##
## Dołączanie pakietu: 'ellipse'
```

```
## Następujący obiekt został zakryty z 'package:graphics':
##
##     pairs
```

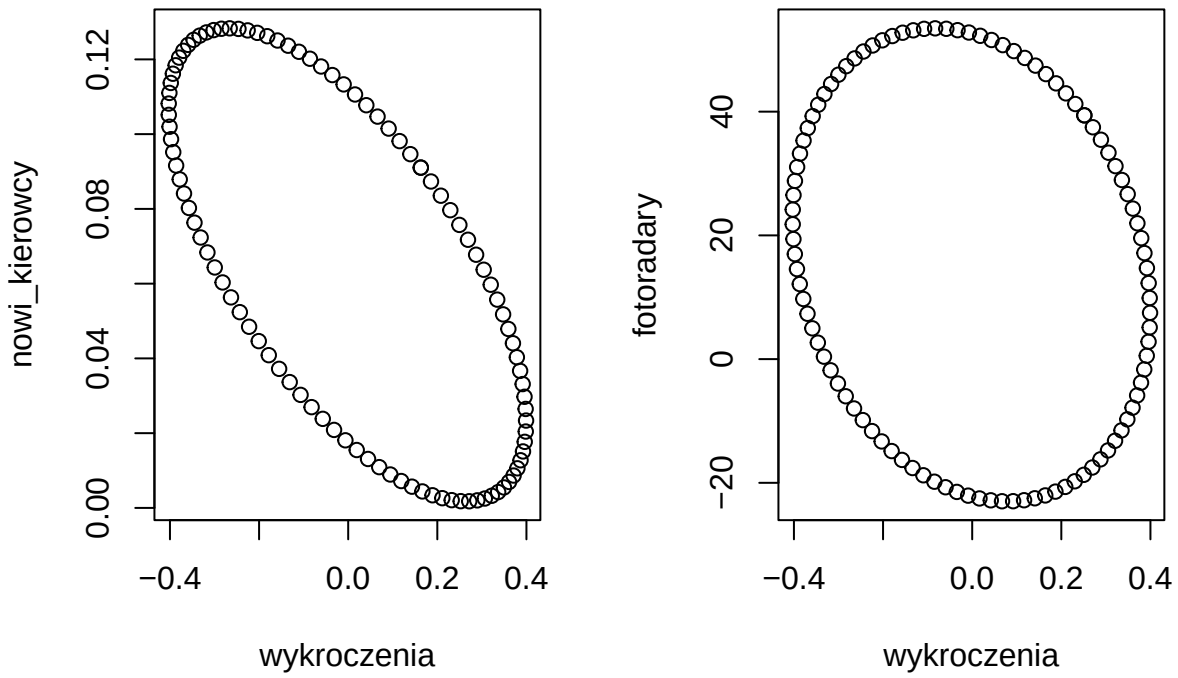



Pierwsza elipsa przedstawia zależność między zmienną wypadki a drogi. Druga elipsa przedstawia zależność między drogi a fotoradary. Obie elipsy przechylone są w prawą stronę co świadczy że ich współczynnik korelacji jest dodatni. Druga elipsa jest bardziej wydłużona, co może sugerować silniejszą zależność między danymi, co za tym idzie w pierwszej elipsie mamy słabszą zależność między danymi.




Obie elipsy przechylone są w lewą stronę, ale nieznacznie co świadczy że ich współczynnik korelacji jest ujemny. Jedna i druga elipsa kształtem zbliżona jest do okręgu co za tym idzie mamy słabą zależność między danymi.

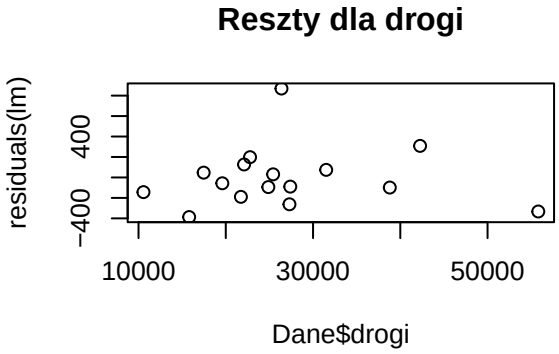



Obie elipsy przechylone są w lewą stronę, co świadczy że ich współczynnik korelacji jest ujemny. Pierwsza elipsa jest bardziej wydłużona, co może sugerować silniejszą zależność między danymi, co za tym idzie w drugiej elipsie mamy słabszą zależność między danymi.

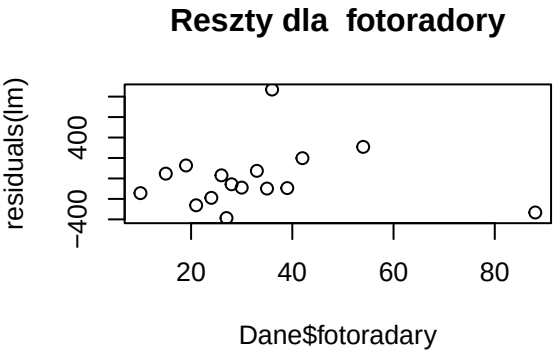
```
##                (Intercept)        drogi  wykroczenia nowi_kierowcy  fotoradary
## (Intercept)     1.00000000  0.091447802  0.030513652   -0.34955860  0.13611984
## drogi           0.09144780  1.000000000  0.009622529   -0.29432027 -0.59924509
## wykroczenia     0.03051365  0.009622529  1.000000000   -0.66208143 -0.20072012
## nowi_kierowcy -0.34955860 -0.294320269 -0.662081429    1.00000000 -0.14869634
## fotoradary      0.13611984 -0.599245093 -0.200720116   -0.14869634  1.00000000
## deszcz         -0.88618684 -0.392026293 -0.099888165    0.42141920  0.01803174
## turysci        -0.17610600  0.253434831 -0.183244418    0.05671721 -0.14451825
##                     deszcz     turysci
## (Intercept)   -0.88618684 -0.17610600
## drogi         -0.39202629  0.25343483
## wykroczenia   -0.09988816 -0.18324442
## nowi_kierowcy  0.42141920  0.05671721
## fotoradary     0.01803174 -0.14451825
## deszcz         1.00000000 -0.09911502
## turysci       -0.09911502  1.00000000
```

Mamy tutaj korelację między danymi egzogenicznymi. Najsłabsza korelacja dodatnia jest pomiędzy drogi-wykroczenia. Najsilniejsza korelacja ujemna jest pomiędzy nowi_kierowcy-wykroczenia. Najsilniejsza korelacja dodatnia jest pomiędzy deszcz-nowi_kierowcy.

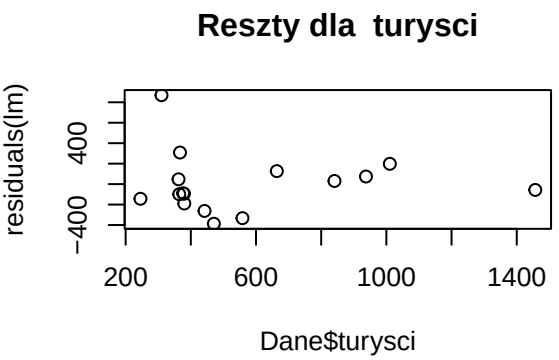
# 4 Analiza reszt pełnego modelu



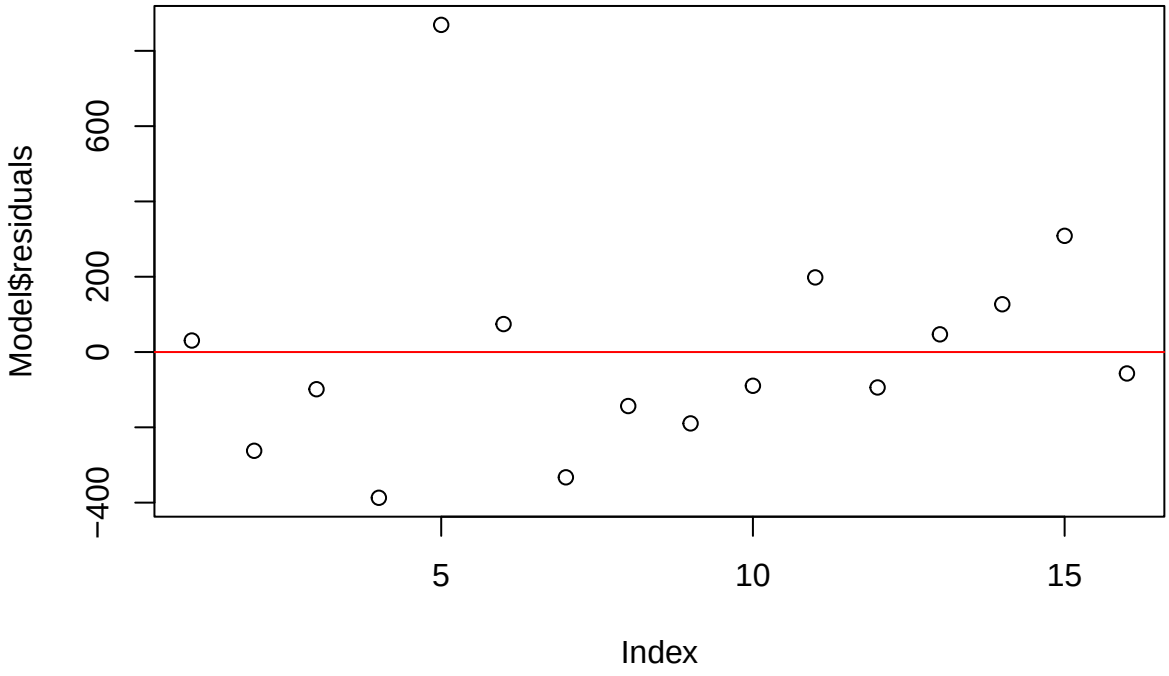


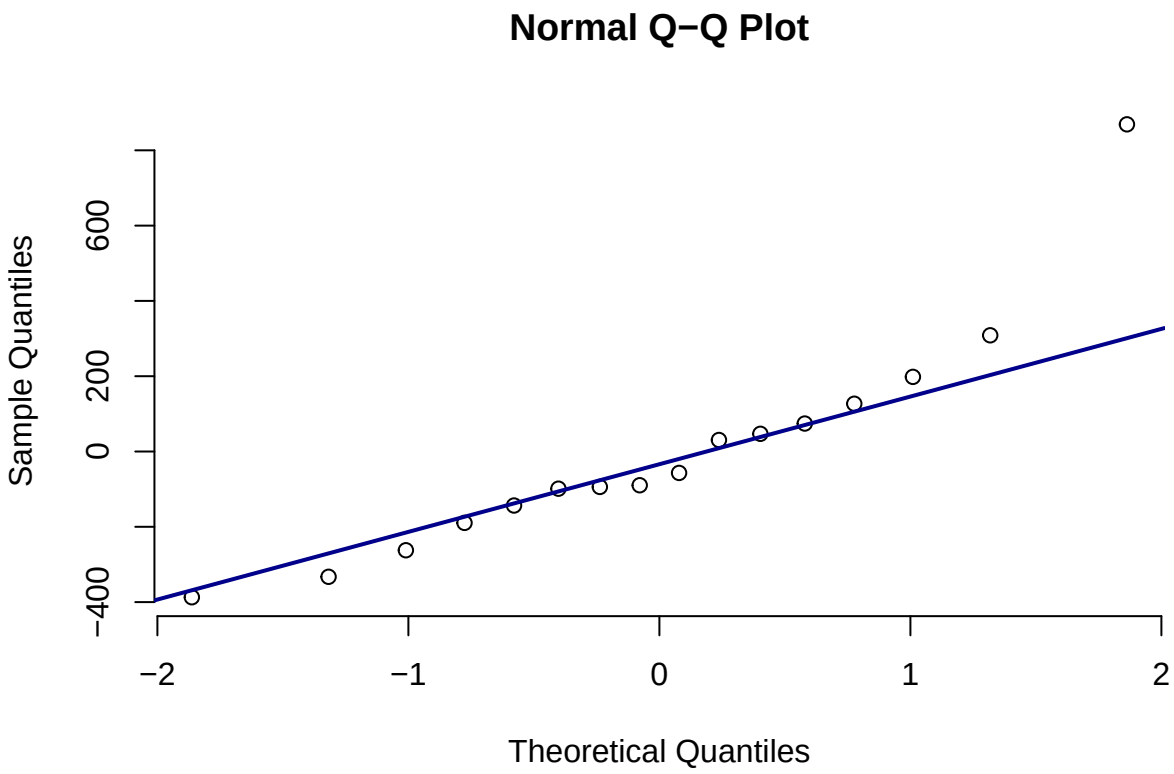



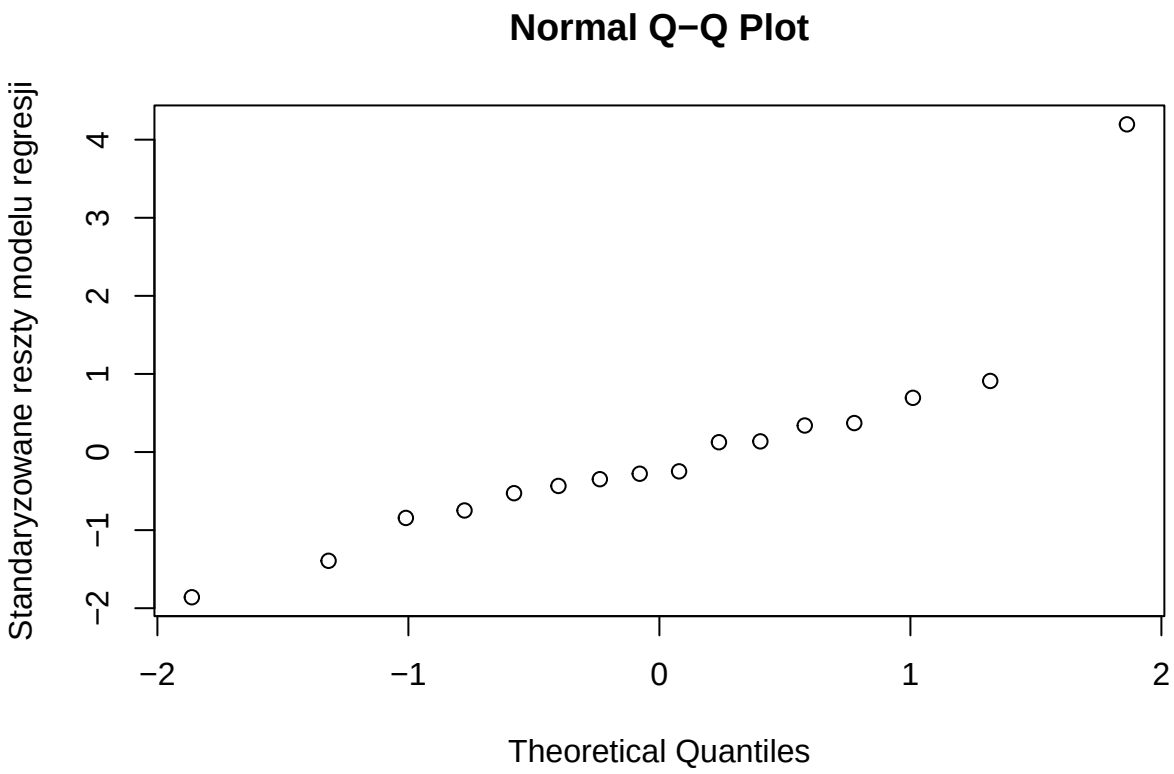




Dla prawie wszystkich wykresów reszty wydają się być rozmieszczone w sposób losowy, chociaż dla zmiennej deszcz i turysci skupiają się bardziej w jednym miejscu. Wykres reszt dla naszego pełnego modelu wygląda następująco:




Normal Q−Q Plot
Sample Quantiles
−400
0
200
600
−2
−1
0
1
2
Theoretical Quantiles




Reszty oscylują wokół 0, ale mamy jedną wartość odstającą. Wskazują na to 3 powyższe wykresy.

Przejdźmy teraz do testu Shapiro-Wilka w celu zbadania normalności.

```
##
##  Shapiro-Wilk normality test
##
## data:  Model$residuals
## W = 0.87327, p-value = 0.03052
```

Mamy tutaj do czynienia z testem gdzie:

H0: Rozkład danych jest normalny

H1: Rozkład danych nie jest normalny.

p wartość jest mniejsza niż 0,05 co świadczy o tym że możemy odrzucić H0. Może to sugerować że reszty nie są zgodne z rozkładem normalnym.

```
##
##  Runs Test
##
## data:  Model$residuals
## statistic = -1.0351, runs = 7, n1 = 8, n2 = 8, n = 16, p-value = 0.3006
## alternative hypothesis: nonrandomness
```

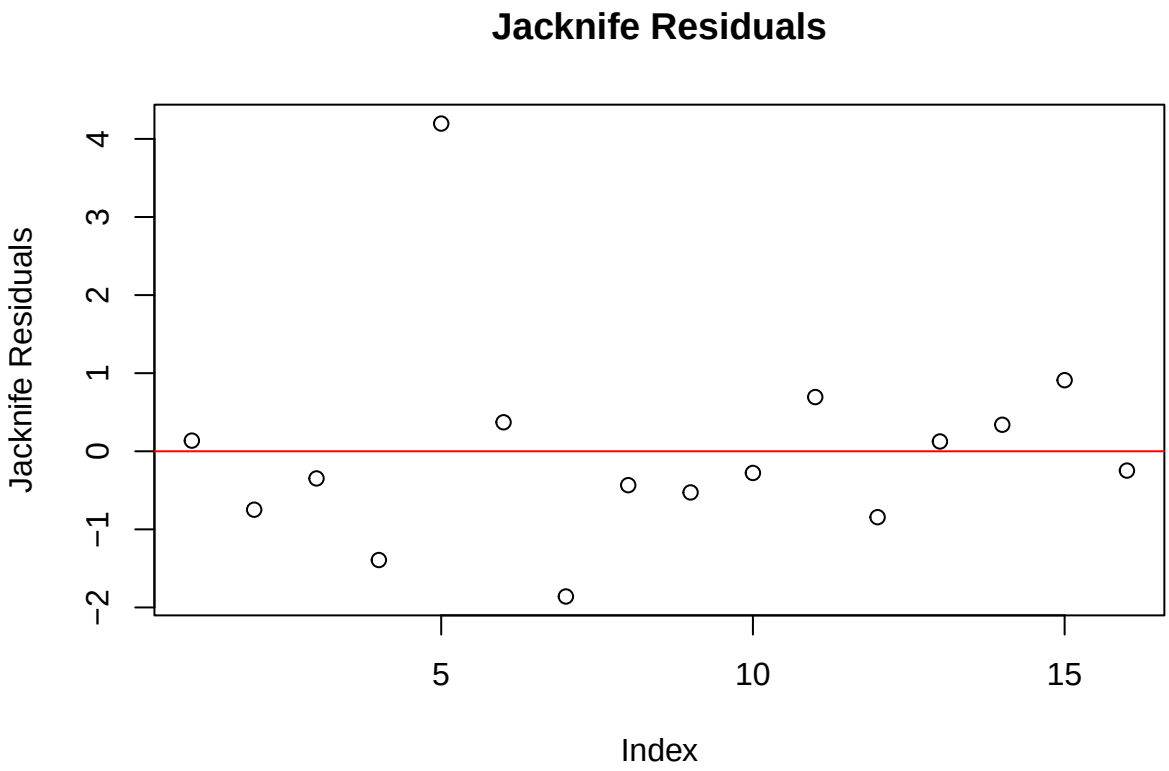
Mamy tutaj do czynienia z testem gdzie:

H0: Sekwencja reszt jest przypadkowa

H1: Sekwencja reszt nie jest przypadkowa.

Wartośc p jest większa od 0,05 więc możemy odrzucić H0, a zatem rozkład reszt nie jest przypadkowy.

W celu dokładniejszej analizy przejdźmy do analizy obserwacji odstających.




Wykres reszt pokazuje jak każda obserwacja wpływa na model po jej usunięciu. Reszty są rozproszone wokół 0, ale jest kilka obserwacji odstających. Obserwacja[5] ma zatem istotny wpływ na model.

```
##        5
## 4.196344
```

Obserwacja która ma największy bezwzględny wpływ na model to obserwacja[5]. Jej duża wartość sugeruje, że ma duży wpływ na wynik modelu.

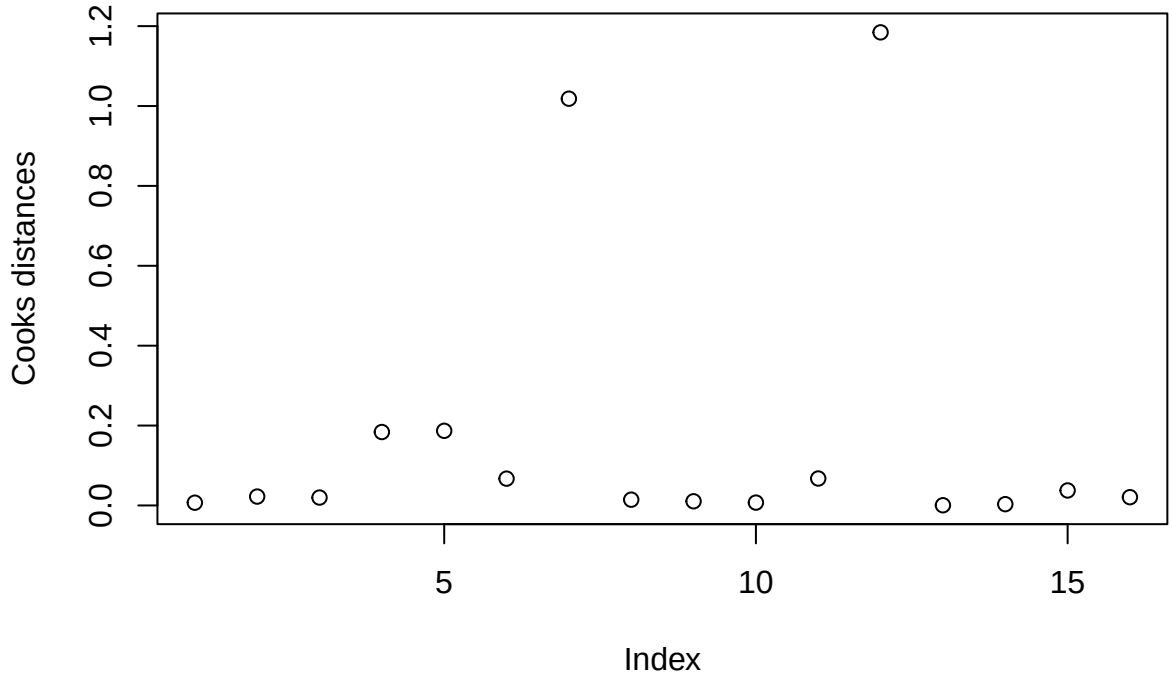
```
## [1] -5.32592
```

Wartość rozkładu t-Studenta dla przedziału ufności 95% ma wartość -5,32 co sugeruje, że wartość t-statystyki jest bardzo mała i znacznie oddalona od zera. Może to wskazywać na znaczącą różnicę między szacowanym współczynnikiem a zerem.

```
##     5
## FALSE
```

Sprawdzamy czy wartość bezwzględna obserwacji odstających jest mniejsza niż lewostrony kwantyl t-rozkładu. Wyszedł nam fałsz, zatem wartość bezwględna obserwacji jest większa niż lewostronny kwantyl t-rozkładu. Więc nasza obserwacja nie jest istotna statystycznie na danym poziomie istnotności.

Przejdźmy teraz do obserwacji wpływowych:




Widać, że dla obserwacji 7 oraz 12 odpowiadające im wartości mają znaczący wpływ na wynik modelu.

```
##            1            2            3            4            5            6
## 0.0068615186 0.0222360858 0.0197485795 0.1838082281 0.1868285952 0.0670814016
##            7            8            9           10           11           12
## 1.0183513823 0.0143871538 0.0104272118 0.0071554544 0.0673884083 1.1844446702
##           13           14           15           16
## 0.0004711185 0.0031723320 0.0375111025 0.0206651921
```

Obserwacje 1,2,3,8,9,10,13,14,15 i 16 mają wartości poniżej 0,05 zatem może to sugerować ich niewielki wpływ na model. Obserwacje 4,5,6 mają umiarkowany wpływ. Natomiast obserwacje 7,11,12 mają znaczny wpływ.

# 5 Analiza regresji

Budujemy teraz model zredukowany regresji ilości wypadków zbudowany na danych, gdzie obserwacje mają odległość cooka < niż największa odległość cooka w całym zestawie danych.

```
##
## Call:
```

```
## lm(formula = wypadki ~ ., data = Dane, subset = (cook < max(cook)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -378.62 -198.54  -63.64  125.10  781.54
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    591.18925 1731.85693   0.341   0.7416
## drogi           -0.04385    0.02642  -1.660   0.1355
## wykroczenia      0.09435    0.18032   0.523   0.6150
## nowi_kierowcy    0.07020    0.02287   3.069   0.0154 *
## fotoradary      13.29910   13.51313   0.984   0.3539
## deszcz          -1.47414    9.95423  -0.148   0.8859
## turysci         -0.16355    0.40569  -0.403   0.6974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 390.8 on 8 degrees of freedom
## Multiple R-squared:  0.8717, Adjusted R-squared:  0.7754
## F-statistic: 9.058 on 6 and 8 DF,  p-value: 0.003276
```

W porównaniu z wyjściowym modelem wartość $R^2$ wyszło większe, wzrosło o 0,26%. Wartość $R^2$ poprawiona spadła o 0,65%. Wartość p wzrosła o 0,001762. W porównaniu do wyjściowego modelu poza zmiennymi wykroczenia i turysci kolejną nieistotną zmienną wydaje się być deszcz.

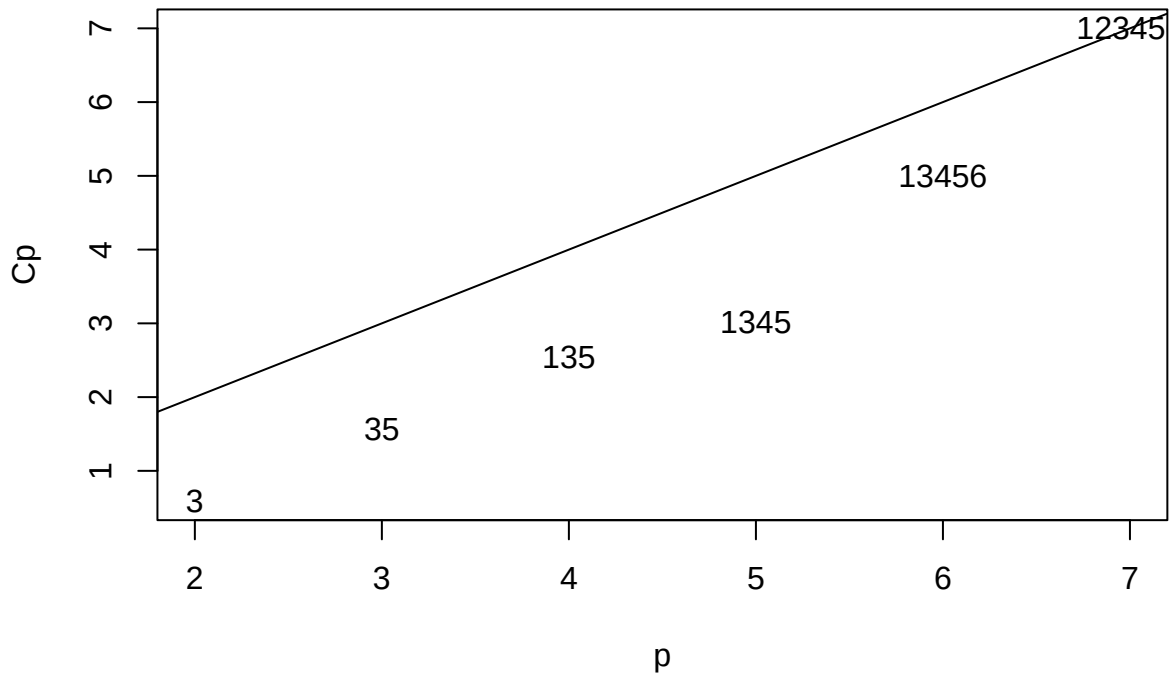
# 6 Kryterium AIC, Ra i Cp

Tworzymy teraz alternatywny model regresji, eliminując obserwacje o największej odległości cooka. Chcemy znaleźć najlepszy model, w pierwszej kolejności sprawdzimy kryteria dla całego modelu i modelu po usunięciu jednej zmiennej.

```
##                df      AIC
## l0              8 242.6581
## ldrogi          7 244.0525
## lwykroczenia    7 240.6585
## lnowi_kierowcy  7 251.7433
## lfotoradary     7 242.8978
## ldeszcz         7 243.4622
## lturysci        7 240.6922
```

W kryterium AIC mała wartość AIC sugeruje, który model jest lepszy, czym jego wartość mniejsza tym model lepszy. Zatem w naszym przypadku najlepszy wydaje się model lwykroczenia

(drogi+nowi_kierowcy+fotoradary+deszcz+turysci).



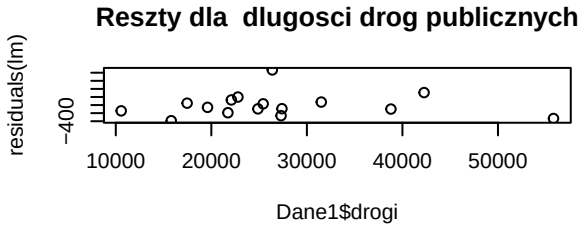

Preferujemy model z najniższym p i $C_p$ , ale analizujemy model tylko bez jednej zmiennej. Zatem lepszy wydaje się model który zawiera zmienne drogi+nowi_kierowcy+fotoradary+deszcz+turysci.

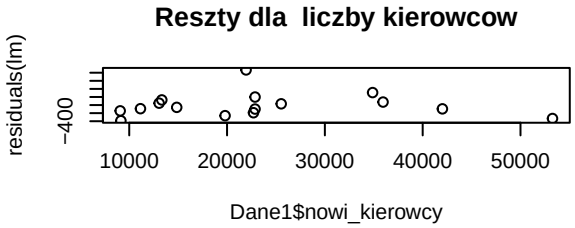
Sprawdźmy teraz informacje o najlepszych modelach na podstawie $R^2$.

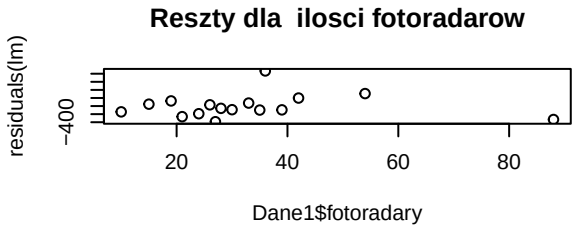
```
##     1,3,4,5       1,3,5         3,5           3   1,3,4,5,6 1,2,3,4,5,6
##       0.821       0.808       0.806       0.804       0.804       0.782
```

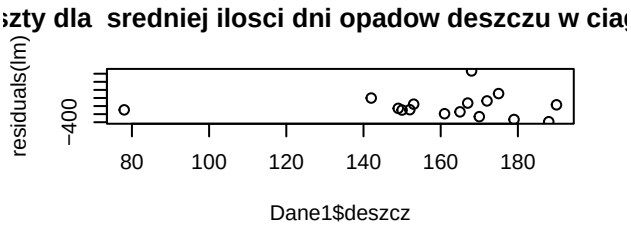
Zgodnie z kryterium R^2 najlepszy model bez wykroczeń i turystów.

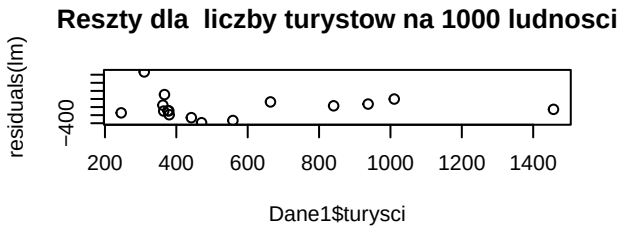
Przejdźmy teraz do analizy reszt modelu bez wykroczeń:



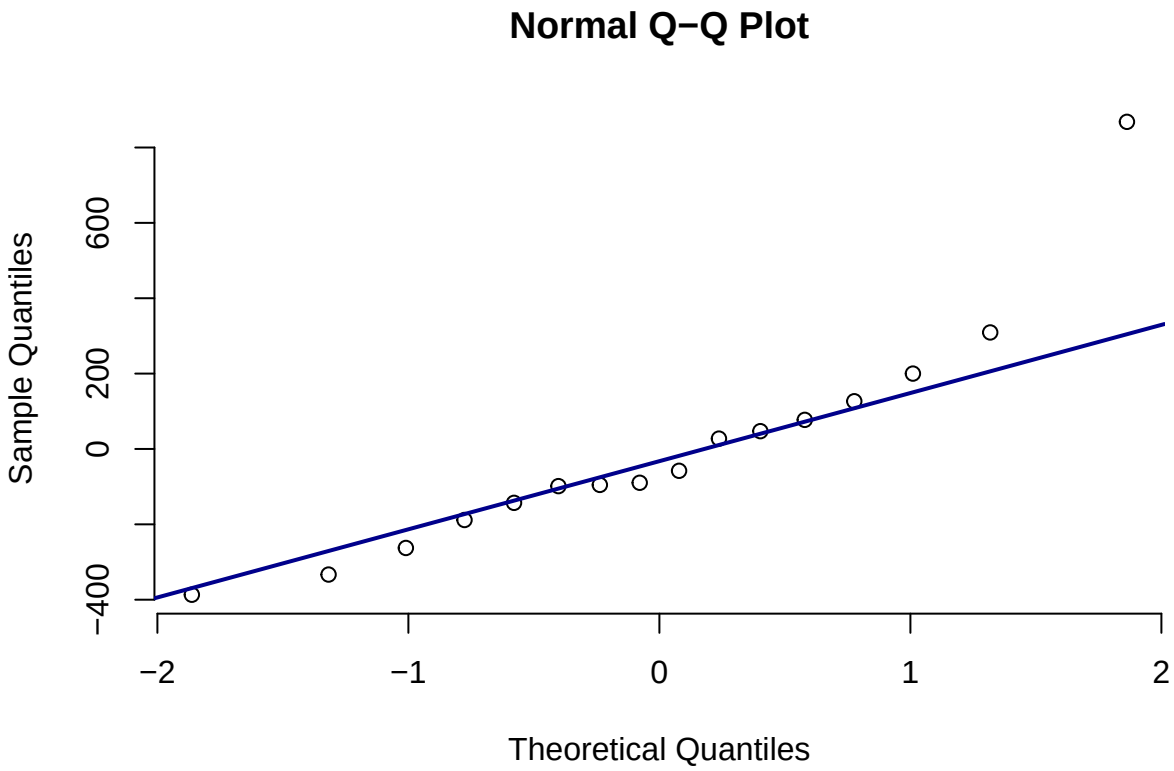



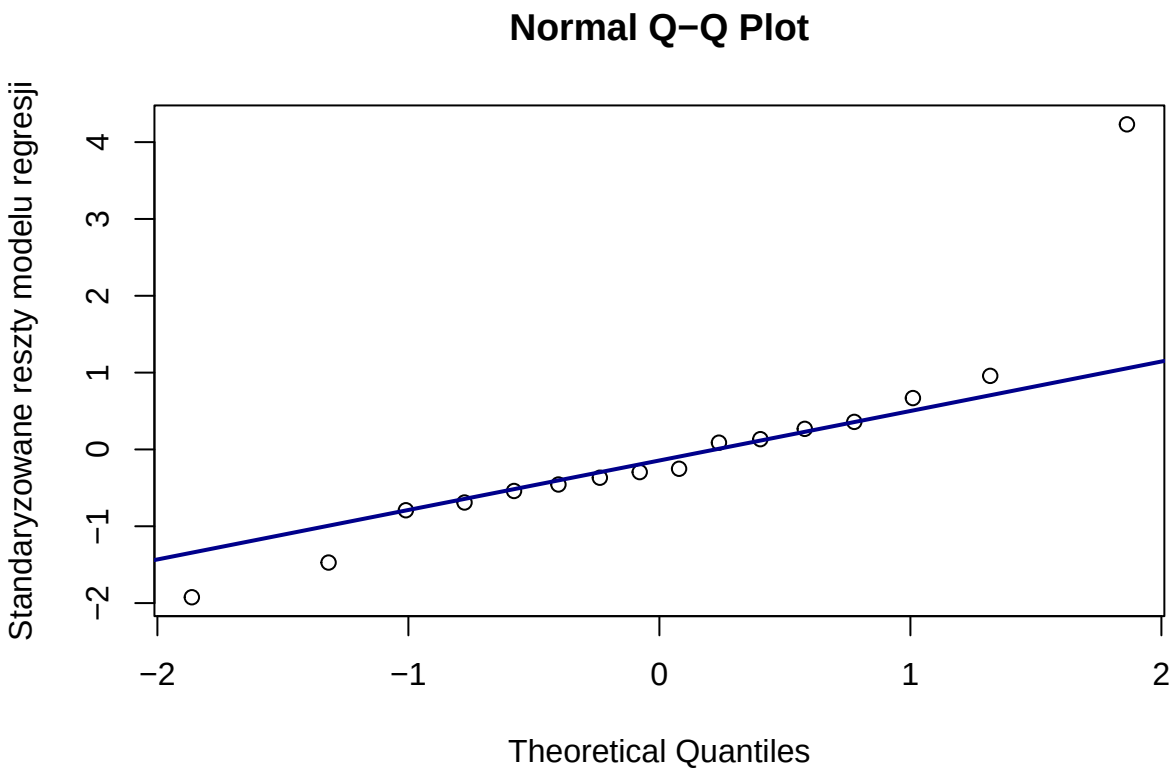

Widać że reszty w modelu po usunięciu zmiennej wykroczenia układają się w mniej losowy sposób.

Nasz model bez wykroczeń:

```
##
## Call:
## lm(formula = wypadki ~ ., data = Dane1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -386.83 -154.39  -73.81   89.58  868.12
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -725.40039  700.38899  -1.036  0.32473
## drogi           -0.03114    0.02026  -1.537  0.15526
## nowi_kierowcy    0.06484    0.01542   4.205  0.00181 **
## fotoradary      15.20008   12.17678   1.248  0.24036
## deszcz           5.96018    4.28936   1.390  0.19483
## turysci          0.04320    0.29604   0.146  0.88688
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 364.7 on 10 degrees of freedom
## Multiple R-squared:  0.8691, Adjusted R-squared:  0.8037
## F-statistic: 13.28 on 5 and 10 DF,  p-value: 0.000379
```

W porównaniu do wyjściowego modelu wzrosła wartość $R^2$ poprawione. Jednak dalej nieistotna wydaje się być zmienna turyści.






Widzimy że reszty dalej oscylują wokół prostej, ale nie usunęliśmy zmiennej odstającej.

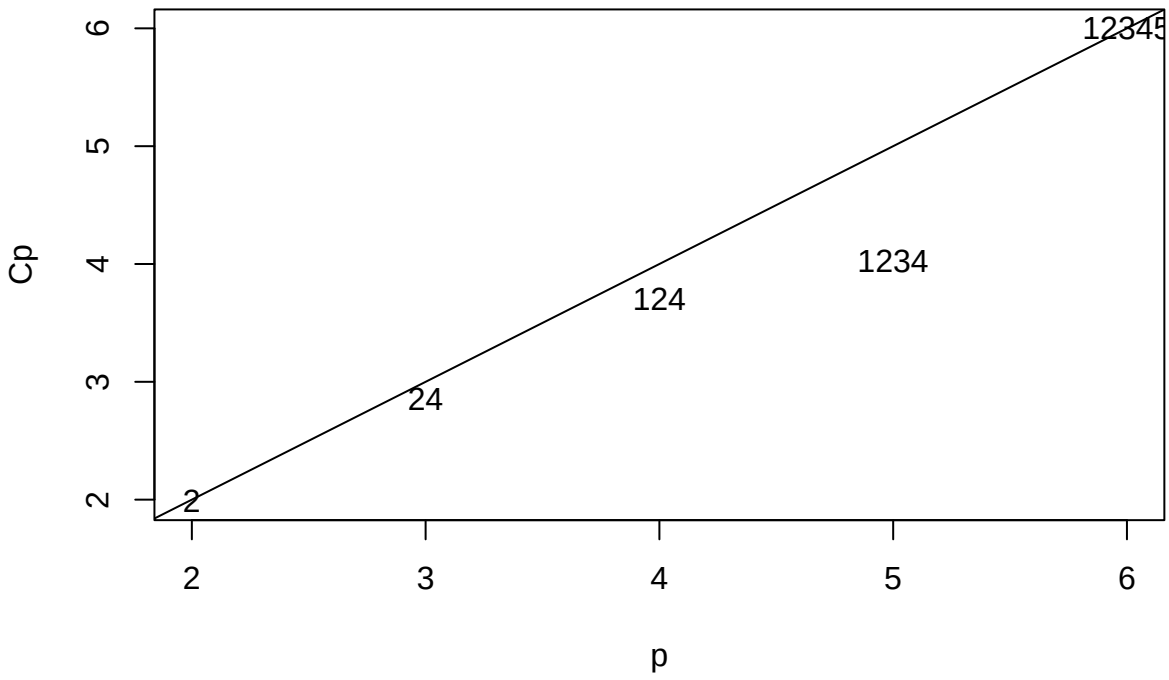
```
##
##  Shapiro-Wilk normality test
##
## data:  Model1$residuals
## W = 0.87382, p-value = 0.03113
```

p wartość jest mniejsza niż 0,05 co świadczy o tym że możemy odrzucić H0. Może to sugerować że reszty nie są zgodne z rozkładem normalnym.

Spróbujmy teraz usunąć z modelu oprócz zmiennej wykroczenia jeszcze jedną zmienną.

```
##                 df      AIC
## ll0              7 240.6585
## lldrogi          6 242.0525
## llnowi_kierowcy  6 254.9520
## llfotoradary     6 240.9754
## lldeszcz         6 241.4830
## llturysci        6 238.6925
```

Widzimy, że z kryterium AIC lepszy model to model llturyści zawierajacy drogi+nowi_kierowcy+fotoradary+deszcz.

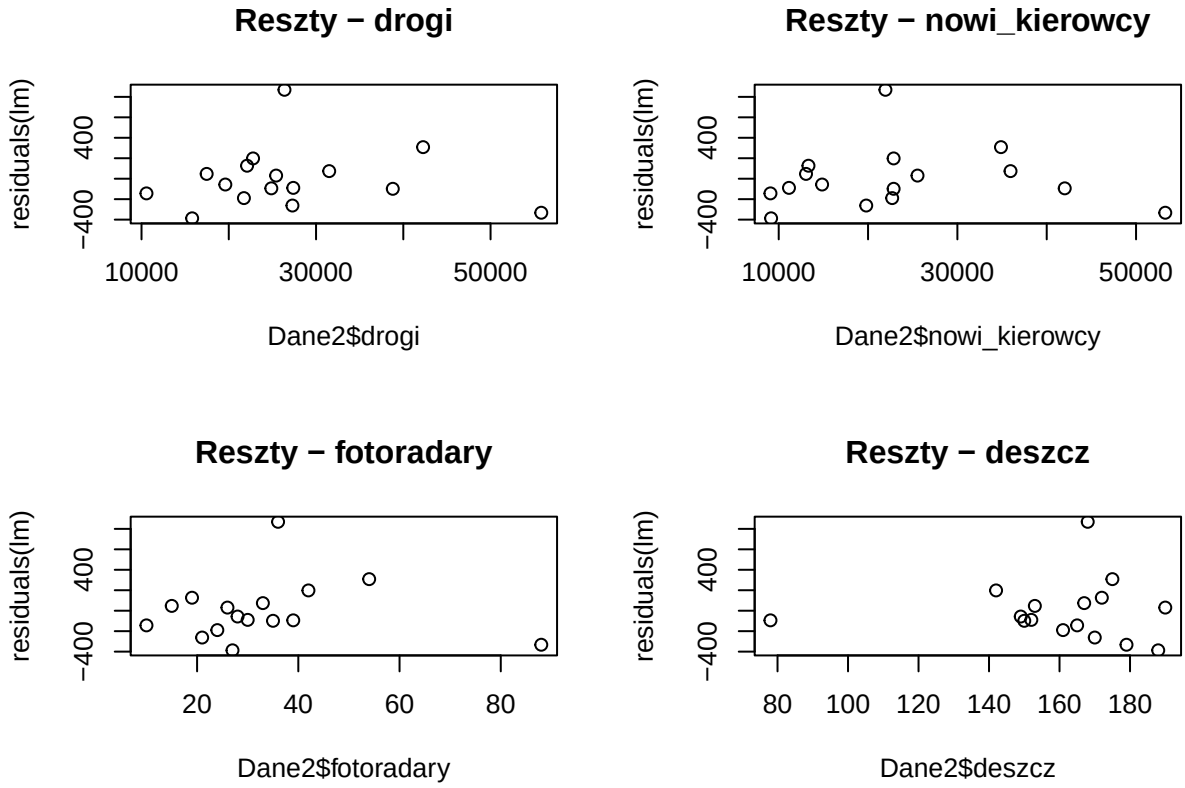



Z kryterium Cp lepszy model to model zawierający drogi+nowi_kierowcy+fotoradary+deszcz.

```
##   1,2,3,4     1,2,4       2,4         2 1,2,3,4,5
##     0.821     0.808     0.806     0.804     0.804
```

Z kryterium Rp najlepszy model to model zawierający drogi+nowi_kierowcy+fotoradary+deszcz.

Zatem trzy kryteria dały jednoznaczną odpowiedź i najlepszy model to model zawierający zmienne drogi+nowi_kierowcy+fotoradary+deszcz.



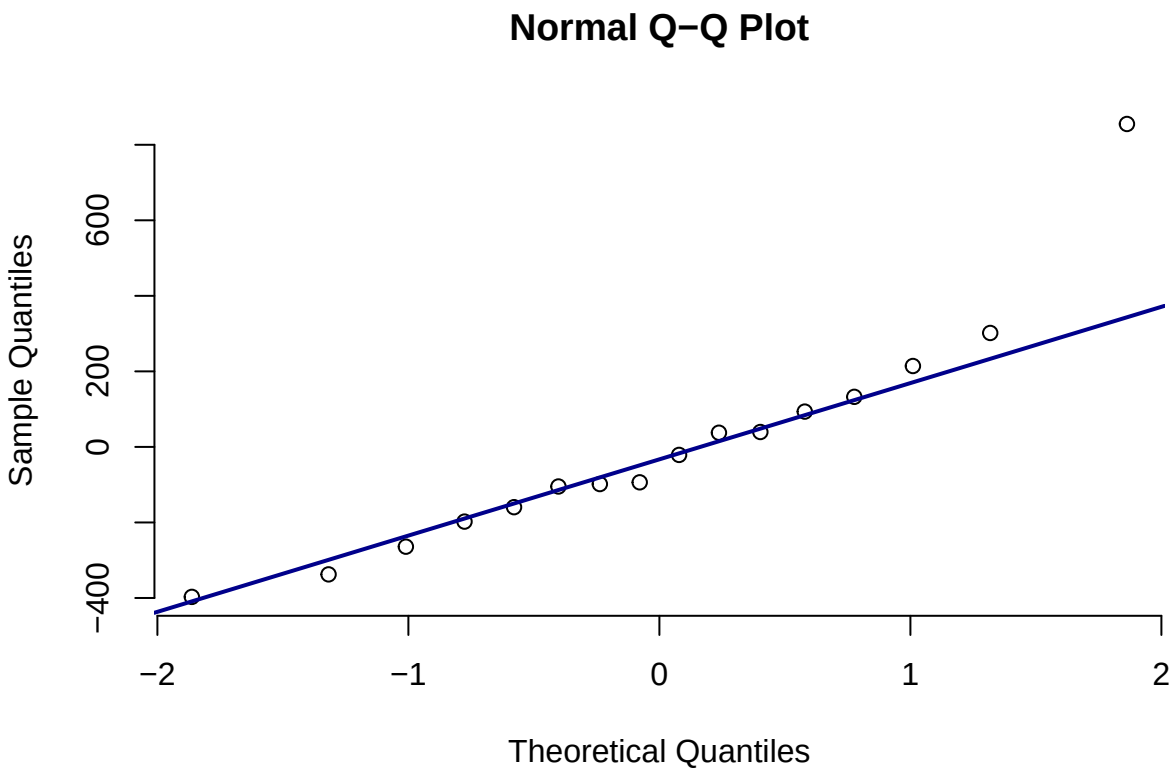

Reszty wydają się być rozmieszczone w mniej losowy sposób, co może sugerować normalność.

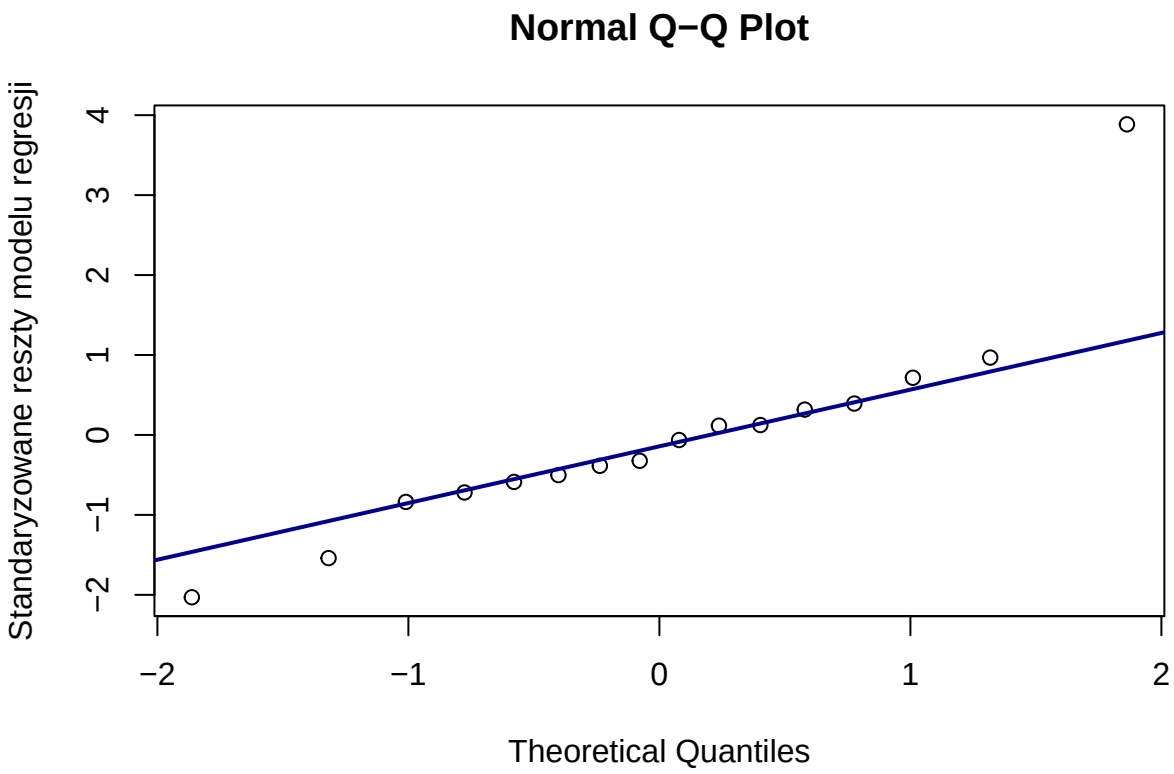
Nasz model bez zmiennej wykroczenia i turyści:

```
##
## Call:
## lm(formula = wypadki ~ ., data = Dane2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -397.07 -169.11  -57.49  103.27  855.04
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -707.66515  658.36284  -1.075    0.305
## drogi           -0.03191    0.01867  -1.709    0.116
## nowi_kierowcy    0.06504    0.01466   4.437    0.001 **
## fotoradary      15.53458   11.41466   1.361    0.201
## deszcz           6.03531    4.06449   1.485    0.166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 348.1 on 11 degrees of freedom
## Multiple R-squared:  0.8688, Adjusted R-squared:  0.8211
## F-statistic: 18.22 on 4 and 11 DF,  p-value: 0.00008124
```

Nasz model ma postać:

Wypadki=-707,67-0.03*drogi+0,07*nowi_kierowcy+15,53*fotoradary+6,04*deszcz.






Podobnie nasze reszty oscylują wokół prostej i mamy jedną wartość odstającą.

```
##
##  Shapiro-Wilk normality test
##
## data:  Model2$residuals
## W = 0.88901, p-value = 0.05373
```

p wartość jest większa niż 0,05 co świadczy o tym że możemy przyjąć H0. Może to sugerować że reszty są zgodne z rozkładem normalnym.

# 7 Analiza końcowa

Najlepszym modelem okazał się model:

Wypadki=-707,67-0.03*drogi+0,07*nowi_kierowcy+15,53*fotoradary+6,04*deszcz.

Największy wpływ na wypadki ma zmienna fotoradar oraz deszcz. Jeśli zwiększymy ilość nowych kierowców o rok, to oczekiwana ilość wypadków zwiększy się o 0,07 jednstek z zasadą cetero paribus.

# 8 Wnioski

Okazało się że największy wpływ na liczbę wypadków względem województw ma ilość fotoradarów oraz średnia ilość dni opadów deszczu w ciągu roku. Jednak istotnym wpływem okazało się być również długość

dróg publicznych oraz liczba nowych kierowców. Nieistotne okazły się być ilość kierowców zatrzymanych oraz ilość turystów na 1000 ludności.