**Część 1.**

**Drzewa decyzyjne do prognozowania deszczu w Australii**

**Część 2.**

**Model regresji nieparametrycznej dla liczby wypożyczeń rowerów w Londynie**

**Część 3.**

**Model MARS dla liczby wypożyczeń rowerów w Londynie**

**Część 1.**

**Drzewa decyzyjne do prognozowania deszczu w Australii**

1. **Opis zbioru danych**

Zbiór danych pochodzi ze strony <https://www.kaggle.com/jsphyg/weather-dataset-rattle-package> i przedstawia dzienne obserwacje z wielu lokalizacji w Australii. Zmienne dotyczą warunków pogodowych. Wszystkie podane temperatury wyrażone są w stopniach Celsjusza.

|  |  |
| --- | --- |
| **Zmienna** | **Opis** |
| MinTemp | Minimalna temperatura w ciągu dnia. |
| MaxTemp | Maksymalna temperatura w ciągu dnia. |
| RainFall | Ilość deszczu w ciągu dnia (w mm). |
| WindGustDir | Kierunek najsilniejszego wiatru (zmienna jakościowa). |
| WindGustSpeed | Prędkość najsilniejszego wiatru (w km/h). |
| WindDir9am | Kierunek wiatru o 9:00 (zmienna jakościowa). |
| WindDir3pm | Kierunek wiatru o 15:00 (zmienna jakościowa). |
| WindSpeed9am | Prędkość wiatru o 9:00 (w km/h). |
| WindSpeed3pm | Prędkość wiatru o 15:00 (w km/h). |
| Humidity9am | Wilgotność powietrza o 9:00 (w procentach). |
| Humidity3pm | Wilgotność powietrza o 15:00 (w procentach). |
| Temp9am | Temperatura o 9:00. |
| Temp3pm | Temperatura o 15:00. |
| Month | Miesiąc. |
| RainTomorrow | Zmienna binarna (1 jeżeli padało następnego dnia, 0 jeżeli nie padało). |

Zmienna RainTomorrow jest zmienną objaśnianą modelowaną w kolejnych częściach pracy. Model ma za zadanie prognozować czy kolejnego dnia będzie padać deszcz, biorąc pod uwagę dane z dnia poprzedniego takie jak temperatura, kierunek i prędkość wiatru oraz wilgotność powietrza. Zbiór danych zawiera oryginalnie około 145 000 obserwacji z lat 2008-2017, jednak ze względu na usunięcie kilku zmiennych (nie opisanych w tabeli powyżej, oryginalnie w zbiorze są 23 zmienne), a następnie usunięciu obserwacji wciąż zawierających braki, pozostało 121 790 obserwacji.

1. **Podstawowa eksploracja danych**

Poniższe wykresy oraz opisy przedstawiają rozkład 4 zmiennych. Jest to zmienna objaśniana oraz 3 zmienne objaśniające, które okazały się najbardziej istotne dla modelu w kolejnych częściach badania.

**Rysunek 1.** Histogram zmiennej objaśnianej RainTomorrow

Chart, histogram

Description automatically generated

Zmienna objaśniająca znacznie częściej przyjmuje wartość 0 – oznaczającą brak deszczu następnego dnia (94 906 obserwacji), w porównaniu do wartości 1 – deszcz następnego dnia (26 884 obserwacji).

**Rysunek 2.** Histogram wilgotności powietrza o 15:00 dnia poprzedniego

Chart, histogram

Description automatically generated

Wilgotność powietrza o godzinie 15:00 zdaje się mieć rozkład podobny do normalnego, jednak ma więcej obserwacji po lewej stronie rozkładu, nie są to duże odchylenia o czym świadczy średnia i mediana w pobliżu środka zakresu wartości. Zakres wartości to (0, 100), średnia wilgotność o 15:00 wynosiła 50,8%, a jej mediana wynosiła 51%.

**Rysunek 3.** Histogram wilgotności powietrza o 09:00 dnia poprzedniego

Chart, histogram

Description automatically generated

Rozkład wilgotności powietrza jest asymetryczny lewostronnie, co oznacza, że większość obserwacji przyjmuje wartość powyżej przeciętnej. Zakres tej zmiennej to także (0, 100), jej średnia wynosi 67,6%, a mediana 69%.

**Rysunek 4.** Histogram najwyższej prędkości wiatru dnia poprzedniego

Chart, histogram

Description automatically generated

Rozkład obserwacji najwyższej prędkości wiatru jest zbliżony do normalnego, dla którego ponownie mediana i średnia arytmetyczna przyjmują podobne wartości. Jej zakres to (7, 135), mediana wynosi 39 km/h, a średnia 40,69 km/h.

1. **Tworzenie drzew decyzyjnych**

Obserwacje zostały w sposób losowy podzielone na podzbiór uczący i testowy, gdzie zbiór testowy to 1/3 wszystkich obserwacji. Pierwsze stworzone drzewo decyzyjne ‘bazowe’, to takie z domyślnymi ustawieniami parametrów w języku R. Drzewo nie zostało sprawdzone krzyżowo z danymi testowymi, ale interpretowalność tego modelu jest wysoka, a jego błąd klasyfikacji na zbiorze testowym wynosi około 0,166.

**Rysunek 5.** Bazowe drzewo decyzyjne

Diagram

Description automatically generated

Wykorzystane zostały jedynie dwie zmienne objaśniające – wilgotność powietrza o 15:00 dnia poprzedniego oraz maksymalna prędkość wiatru dnia poprzedniego. Najważniejszą zmienną jest wilgotność powietrza o 15, która już przy pierwszym węźle drzewa klasyfikuje zmienną objaśnianą jako 0, gdy wilgotność wynosiła mniej niż 72%. Aż 84% wszystkich obserwacji znalazło się z tym liściu, daje to jedynie 14% prawdopodobieństwa na deszcz pomimo niskiej wilgotności powietrza. Jeżeli wilgotność powietrza jest wyższa niż 82%, obserwacje klasyfikowane są jako 1 (nastąpi deszcz następnego dnia), a prawdopodobieństwo na to wynosi 0,8. Jeżeli wilgotność powietrza znajduje się w przedziale pomiędzy 72% a 82%, kolejną zmienną różnicującą będzie maksymalna prędkość wiatru – jeżeli jest ona niższa niż 42 km/h, model prognozuje brak deszczu, jeżeli jest wyższa model prognozuje deszcz następnego dnia. Należy jednak zauważyć, że dla tych dwóch liści prognozowane prawdopodobieństwo jest bliższe 0,5, niż w przypadku obserwacji, których wilgotność powietrza jest niższa niż 72%, lub wyższa niż 82%.

Następnie obliczone zostało pełne drzewo decyzyjne, które ze względu na bardzo dużą wielkość nie może zostać zwizualizowane, nie przynosi ono także najlepszych wyników ze względu na zbyt duże dopasowanie do obserwacji w próbie uczącej.

**Rysunek 6.** Wykres relatywnego błędu drzewa w sprawdzaniu krzyżowym względem wielkości drzewa

Chart, histogram

Description automatically generated

Na podstawie rysunku 6. można stwierdzić, że zwiększanie wielkości drzewa powyżej około 85 węzłów nie mają sensu – ich zdolność predykcyjna dla nowych danych zmniejsza się.

Drzewo zostało przycięte na dwa sposoby – drzewo optymalne, to takie dla którego przyjęte kryterium stopu wynosiło najmniejszy możliwy relatywny błąd sprawdzania krzyżowego dodać trzy jego odchylenia standardowe. Zostało stworzone także drugie drzewo, dla którego relatywny błąd w sprawdzaniu krzyżowym jest najmniejszy (nazywane dalej najlepszym drzewem), jednak jego interpretacja jest trudna z powodu bardzo wielu węzłów.

Na rysunku 7 przedstawione zostało drzewo optymalne. Błąd klasyfikacji na próbie testowej tego drzewa wynosi 0,158, co jest znacznie lepszym wynikiem niż błąd drzewa bazowego opisanego wcześniej, oraz jedynie trochę groszy niż błąd 0,159 drzewa najlepszego. Zaletą drzewa optymalnego jest jego możliwa interpretacja wizualna.

**Rysunek 7.** Optymalne drzewo decyzyjne

Timeline

Description automatically generated

Względem bazowego to drzewo ma znacznie więcej węzłów oraz liści. Wciąż najważniejszym predyktorem jest wilgotność powietrza o 15:00 dnia poprzedniego, który dzieli ponad 60% obserwacji w próbie uczącej. Jeżeli jest on niższy niż 54% to model prognozuje brak deszczu w kolejnym dniu (jest to 55% wszystkich obserwacji, prawdopodobieństwo deszczu wynosi wtedy 8%), a jeżeli jest on wyższy niż 82% to model klasyfikuje obserwacje jako 1 – wystąpi deszcz w dniu kolejnym. Ważnymi predyktorami są także maksymalna prędkość wiatru, kierunek wiatru o godzinie 15, minimalna temperatura, ilość deszczu w dniu poprzedzającym predykcje oraz wilgotność powietrza o 9:00.

Istotność poszczególnych zmiennych na prognozę modelu jest przedstawiona na rysunku 8.

**Rysunek 8.** Istotność zmiennych w optymalnym drzewie decyzyjnym

Table

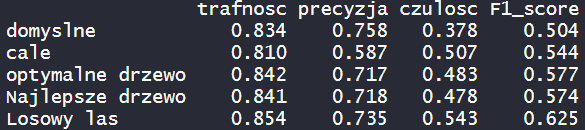
Description automatically generated

Niektóre ze zmiennych mają znikomą istotność, a wilgotność o 15:00 dnia poprzedniego jest znacznie istotniejsza dla predykcji od wszystkich pozostałych zmiennych.

1. **Porównanie modeli oraz wnioski**

Dodatkowym estymowanym modelem dla celu porównania miar predykcji jest model losowego lasu, który został policzony dla 500 drzew oraz 3 możliwych predyktorów wylosowany dla każdego węzła. Wszystkie drzewa zostały porównane ze względu na miary trafności, prezycji, swoistości oraz F1 score.

**Rysunek 9.** Przegląd miar predykcji



Ze względu na rozkład zmiennej objaśnianej – dla której jest znacznie więcej obserwacji 0 – brak deszczu, niż wartości 1 – wystąpienie deszczu, F1 score będzie lepszą miarą predykcji niż trafność. Jest tak ponieważ model częściej prognozujący niewystąpienie deszczu będzie miał dużą trafność oraz precyzje, jednak nie będzie miał własności predykcyjnych i jego czułość będzie niska.

Najlepszym modelem ze względu na kryterium F1 jest model losowego lasu, jednak ze względu na brak interpretowalności i łatwej wizualizacji, model optymalny także ma bardzo dobre miary predykcji.

Każdy z modeli ma dość dużą trafność (około 80-85%) co oznacza, że większość predykcji wszystkich modeli jest poprawna.

Precyzja, która mówi o tym jak dużo jest poprawnych prognoz deszczu wśród wszystkich prognoz wystąpienia deszczu, także jest podobny dla wszystkich modeli (wynosi 70-75%) z wyjątkiem drzewa nieprzyciętego.

Czułość to miara, która daje informacje o tym jak rzadko model przewiduje deszcz gdy go nie ma. Tutaj widać najbardziej rozbieżność pomiędzy modelami, najlepszą wartość ma losowy las, około 6 punktów procentowych więcej niż optymalne drzewo decyzyjne.

Najlepszym modelem predykcji będzie losowy las, a najlepszym modelem to opisu i zgłębienia zmiennych istotnych dla predykcji będzie drzewo optymalne. Należy zauważyć, że czułość dla wszystkich modelów jest dość niska, co oznacza, że często występują błędy typu false-negative – prognoza braku deszczu gdy on występuje, co sugeruje, że może w modelu brakować jakiejś istotnej zmiennej objaśniającej.

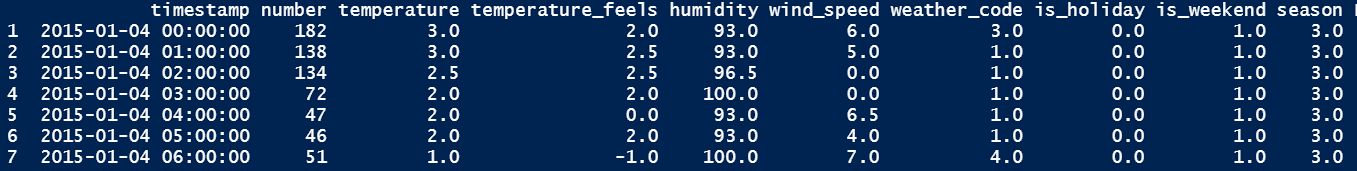
**Część 2. Model regresji nieparametrycznej dla liczby wypożyczeń rowerów w Londynie**

1. **Wstęp**

Dane do tej części pochodzą ze strony <https://www.kaggle.com/hmavrodiev/london-bike-sharing-dataset> i przedstawiają ilość wypożyczeń rowerów w Londynie w okresie od 1 kwietnia 2015r. do 1 marca 2017r. Częstotliwość danych jest tutaj godzinowa, mamy do czynienia z 17413 unikalnymi obserwacjami, każda z nich zawiera liczbę nowych wypożyczeń rowerów w przeciągu danej godziny. Dodatkowo, każdorazowo podana jest tutaj temperatura (oryginalnie *temperature*), temperatura odczuwalna (*temperature\_feels*), wilgotność powietrza (*humidity*), prędkość wiatru *wind\_speed*), pogoda (*weather\_code*), pora roku (*season*) oraz wskazanie, czy jest weekend (*is\_weekend*) jak również, czy jest to dzień świąteczny (*is\_holiday*).

Fragment tabeli z danymi pokazany jest na Rysunku 10.

**Rysunek 10.** Fragment tabeli z danymi źródłowymi

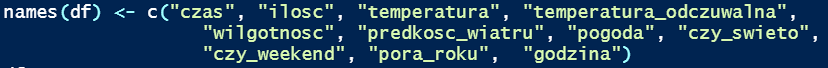


Szczegóły dotyczące oryginalnych zmiennych są następujące:

* *timestamp* – data oraz godzina odnosząca się do ilości wypożyczeni
* *number* – ilość nowych wypożyczeń rowerów
* *temperature* – temperatura podana w stopniach Celsjusza
* *temperature\_feels* – temperatura odczuwalna (w stopniach Celsjusza)
* *humidity*  - wilgotność podana w procentach
* *wind\_speed* – prędkość wiatru w km/h
* *weather\_code* – 7 różnych kategorii pogody zakodowanych następująco: 1 – *clear* (przeważnie bezchmurnie, ewentualnie z lekkim zamgleniem) *,* 2 – *scattered clouds/few clouds* (niewielka ilość chmur bez deszczu), 3 -  *broken clouds* (zwykle wiele małych chmur), 4 – *cloudy* (pochmurno, chmury deszczowe), 7 -  *rain/* *light rain shower/light rain* (deszcz), 10 – *rain with thunderstorm* (deszcz z burzą) i 26 – *snowfall* (opady śniegu)
* *is\_holiday* – 1 jeśli dany dzień jest dniem świątecznym, 0 w przeciwnym przypadku
* *is\_weekend* – 1 jeśli jest to dzień weekendowy, 0 w przeciwnym przypadku
* *season* – pora roku: 0 – wiosna, 1- lato, 2- jesień, 3 – zima.

Oryginalna nazwa bazy danych *london\_merged* została przez nas zmieniona na *df.* Dodatkowo, ze zmiennej *timestamp* wyodrębniliśmy godzinę wypożyczenia (jest ona z zakresu 0 do 23, gdzie 0 oznacza północ – początek danego dnia), zaś wszystkim zmiennym nadaliśmy polskie nazwy, zgodnie z Rysunkiem 11

**Rysunek 11.** Nowe nazwy zmiennych



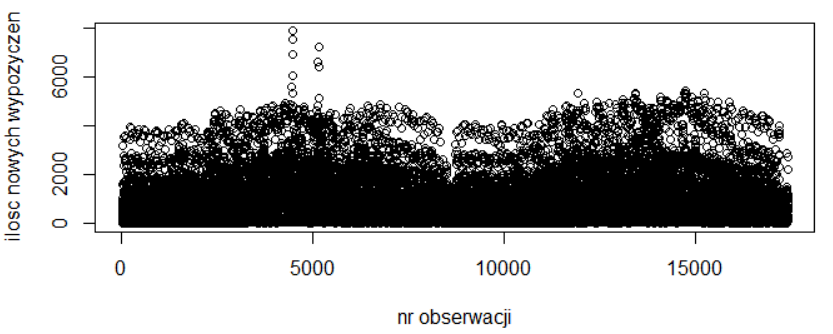
1. **Przegląd danych**

Zacznijmy od podstawowego przeglądu danych, w celu wyeliminowania błędów danych oraz zlokalizowania ewentualnych braków danych.

Ilość wypożyczeń

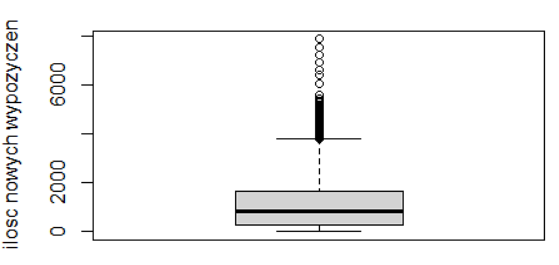
Na Rysunku 12. Przedstawiony jest wykres zależności numeru obserwacji od ilości wypożyczeń.

**Rysunek 12.** Wykres ilości wypożyczeń w zależności od obserwacji

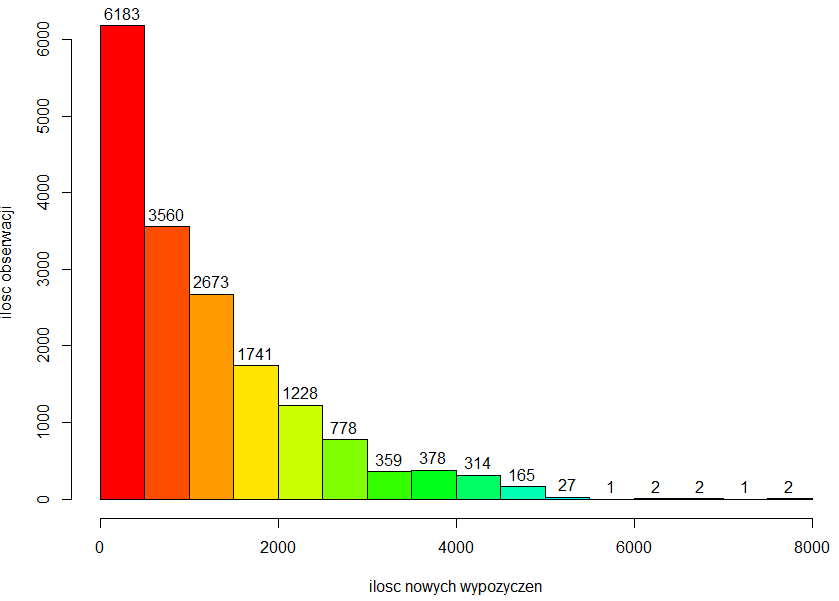


Widzimy, że ilość nowych wypożyczeń rowerów z reguły była nie większa niż 4000. Dodatkowo można zauważyć, że części odpowiadające za pierwszy i drugi rok wypożyczeń są do siebie podobne. Spostrzeżenia co do ilości nowych wypożyczeń potwierdzają się na wykresie pudełkowym przedstawionym na Rysunku 13 oraz na histogramie przedstawionym na Rysunku 14. Maksymalne liczby sięgają około 8000. Są to pomiary podobnego rzędu wielkości, liczby są wiarygodne, więc zostawiamy wszystkie te obserwacje do analizy.

**Rysunek 13.** Wykres pudełkowy ilości nowych wypożyczeń



**Rysunek 14.** Histogram ilości nowych wypożyczeń

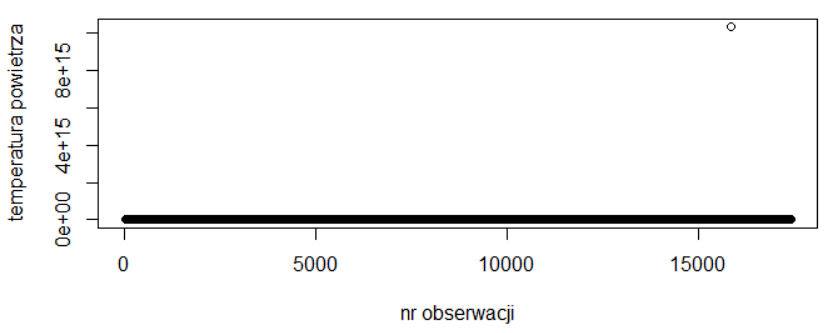


Dodatkowo zgodnie z wyliczeniami programu R, mediana ilości nowych wypożyczeń w rozważanym okresie wynosiła 844 na godzinę, pierwszy kwartyl 257, zaś trzeci 1672.

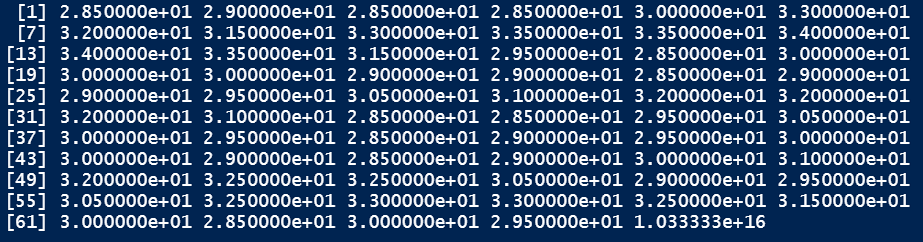
Temperatura

Podobnie jak poprzednio zacznijmy od wykresu temperatur dla wszystkich obserwacji (Rysunek 15.):

**Rysunek 15.** Wykres temperatur (błędna obserwacja)

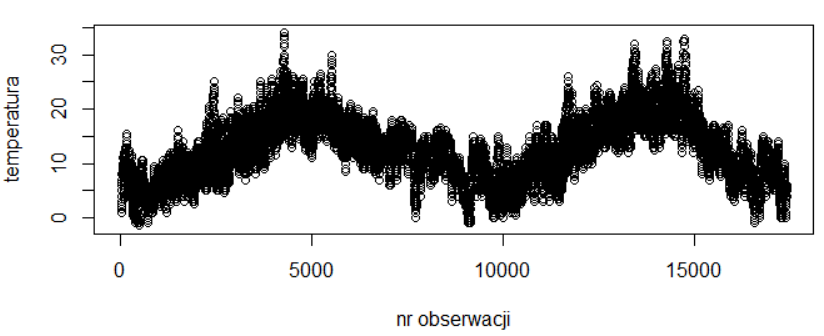


Niestety, z tego wykresu nie dowiemy się, jakie (z jakiego zakresu) temperatury miały miejsce dla obserwacji, ponieważ mamy tutaj do czynienia z ewidentnym błędem w zapisie temperatury w okolicach 16000. obserwacji (jest to temperatura, która wynosi 1,033\*1016 stopni Celsjusza, co jest widoczne po wypisaniu przez program R listy najwyższych temperatur (Rysunek 16.)

**Rysunek 16.** Najwyższe temperatury w zbiorze danych

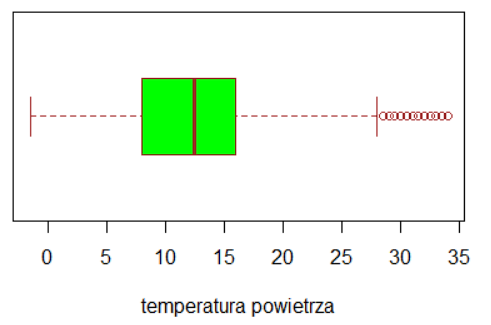
Na podstawie tego spisu, widać że jest to jedyna temperatura błędnie wpisana w zbiorze danych. Pozostałe temperatury, choć relatywnie wysokie, są jednak wiarygodnymi pomiarami, a w czasach globalnego ocieplenia ich wartości tym bardziej są pożądane w celach analizy. Po usunięciu obserwacji odpowiadającej błędnej temperaturze (obserwacja nr 15834), otrzymujemy wykres przedstawiony na Rysunku 17.

**Rysunek 17.** Wykres temperatur po usunięciu błędnego wpisu



Bazując na powyższym wykresie możemy stwierdzić, że temperatury w rozważanym okresie były w znakomitej większości dodatnie i mniejsze niż 30 stopni. Zauważalna jest również oczywista cykliczność ze względu na około dwuletni okres obserwacji. Nieco więcej konkretów dostajemy analizując wykres pudełkowy temperatur (Rysunek 18.).

**Rysunek 18.** Wykres pudełkowy temperatur



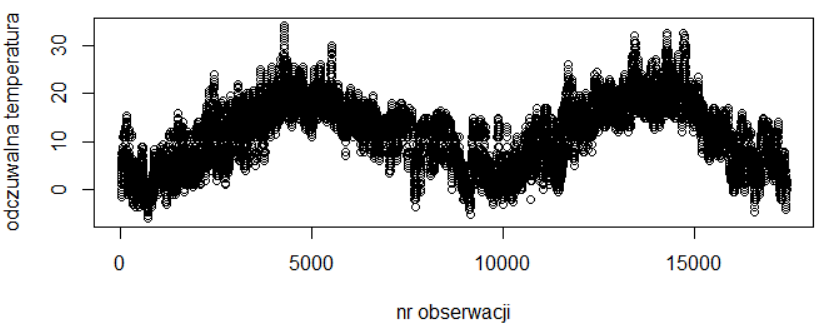
Widzimy, że połowa pomiarów mieści się między około 7-8, a 16 stopniami Celsjusza.

Konkretnie, mediana temperatur wynosi 12,5 stopnia, pierwszy kwartyl 8, zaś trzeci 16 stopni.

Temperatura odczuwalna

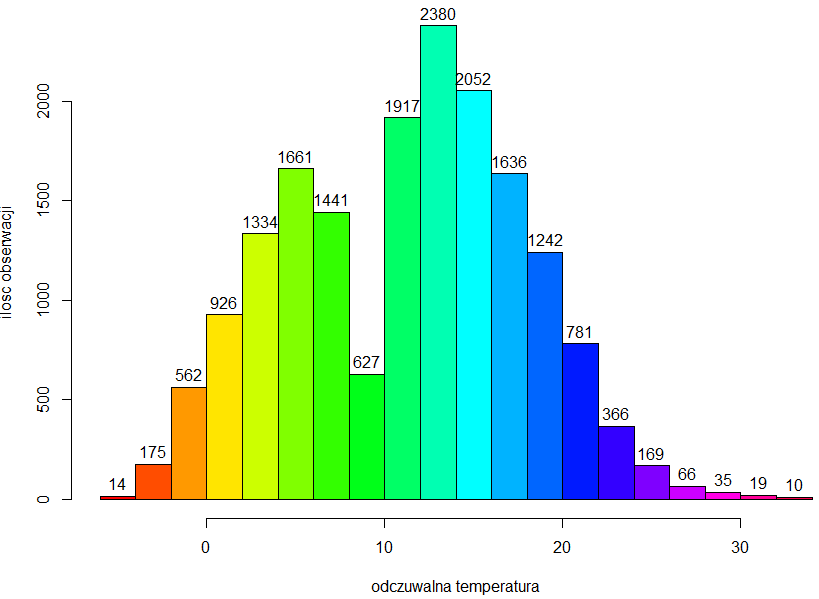
Tutaj sytuacja jest identyczna, jak przy temperaturze powietrza. Ponownie mamy błąd danych przy obserwacji nr 15834. Po jej usunięciu wykres temperatury odczuwalnej przedstawia się następująco (Rysunek 19.):

**Rysunek 19.** Wykres temperatury odczuwalnej



Łatwo zauważyć, że jest to wykres o bliźniaczym przebiegu jak ten z Rysunku 17. Podobnie wygląda również histogram rozkładu temperatur, z pierwszym kwartylem 6, medianą 12,5 i średnią 16 stopni. Przyjrzyjmy się tym razem histogramowi rozkładu odczuwalnych temperatur (Rysunek 20).

**Rysunek 20.** Histogram temperatur odczuwalnych

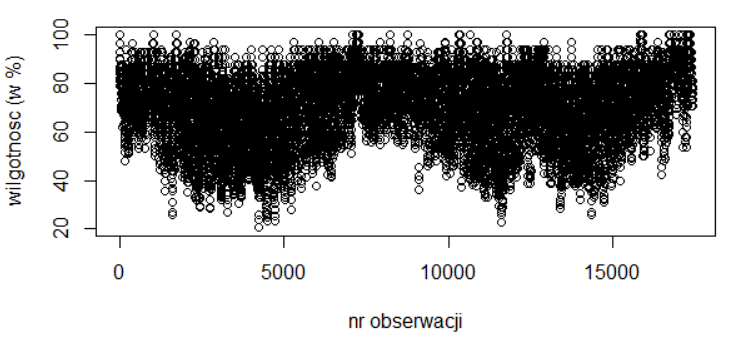


Widzimy, że temperatura odczuwalna w rozważanym okresie rzadko przekraczała 22 stopni Celsjusza, podobnie niewiele było obserwacji z temperaturą odczuwalną mniejszą od zera. Najczęściej temperatura odczuwalna zawierała się między 10 a 16 stopni Celsjusza.

Wilgotność

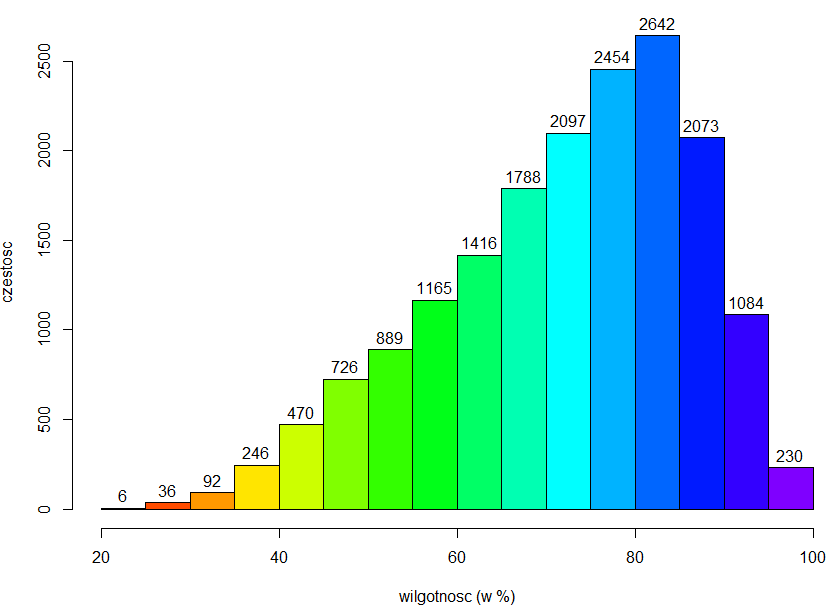
Z uwagi na Rysunek 21. nie stwierdzamy błędów w wartościach tej zmiennej (jest ona wyrażona w procentach).

**Rysunek 21.** Wykres wartości wilgotności w rozważanym okresie



Widzimy, że w analizowanym okresie przeważały godziny o wysokich wartościach wilgotności, zauważalna jest również okresowość w przebiegu tego szeregu czasowego. Wydaje się również, że najwięcej okresów było z wilgotnością w granicach 60-85%. Potwierdzeniem tych spostrzeżeń jest poniższy histogram (Rysunek 22.)

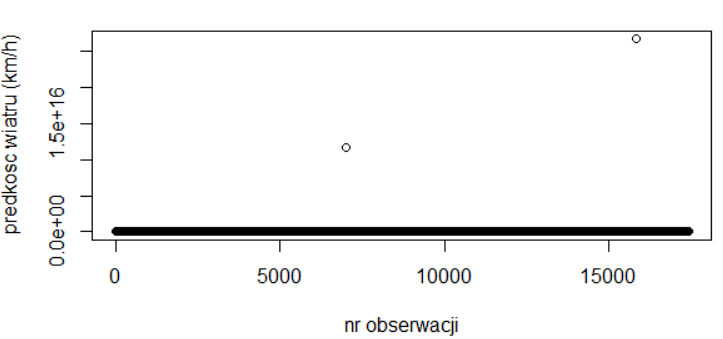
**Rysunek 22.** Histogram wilgotności



Prędkość wiatru

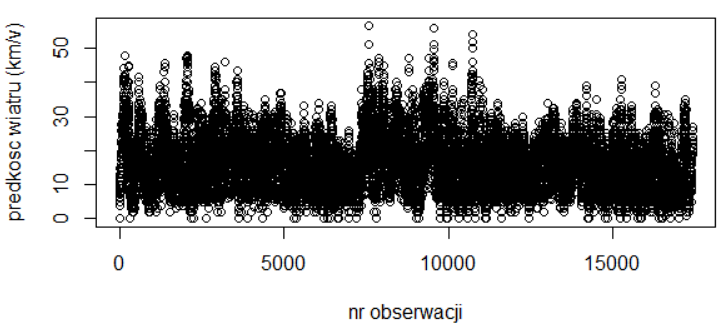
Prędkość wiatru w analizowanym zbiorze danych podawana jest w km/h. Mamy tutaj do czynienia z co najmniej dwoma błędnymi wpisami, o czym świadczy wykres z Rysunku 23.

**Rysunek 23.** Wykres prędkości wiatru (błędne obserwacje)



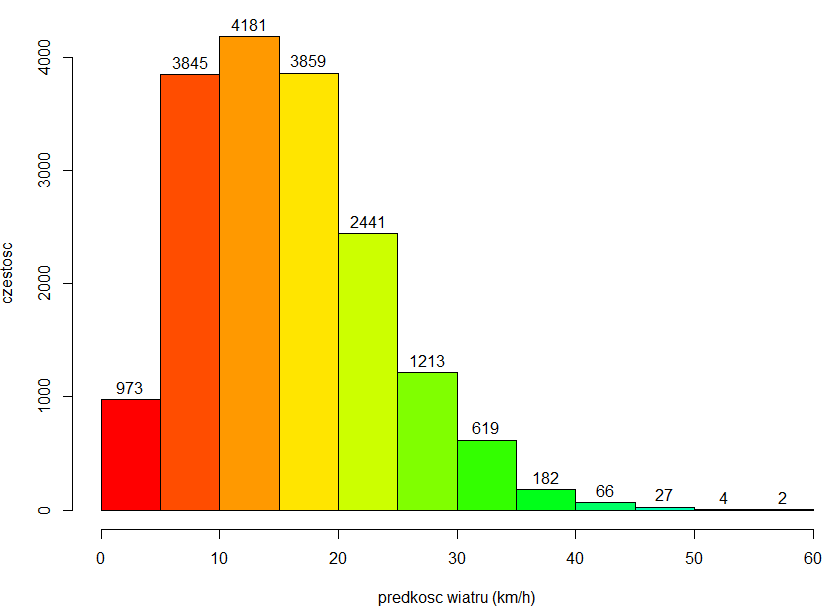
W związku z powyższym z danych wyrzucamy obserwacje nr 7027 i 15834. Ponowna próba narysowania wykresu prędkości wiatru przynosi dobre rezultaty (Rysunek 24.).

**Rysunek 24.** Wykres prędkości wiatru (po odrzuceniu błędnych obserwacji)



Otrzymujemy w pełni wiarygodny wykres, z którego wynika, że prędkość wiatru w przeciągu rozważanych około dwóch lat rzadko przekraczała 40 km/h, zaś najczęściej oscylowała między 10 a 25 km/h. Analizując histogram zaprezentowany na Rysunku 25 widzimy również, że często prędkość ta osiągała wartości między 5 a 10 km/h.

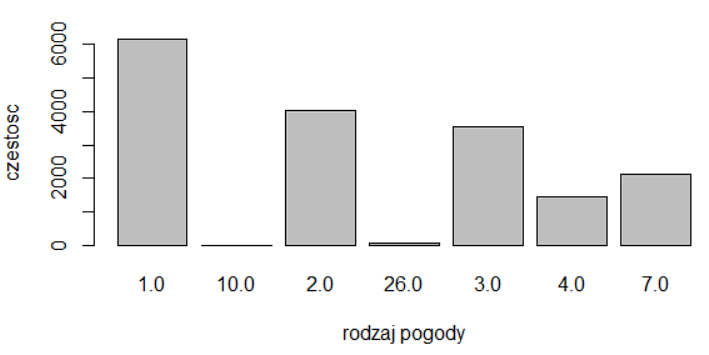
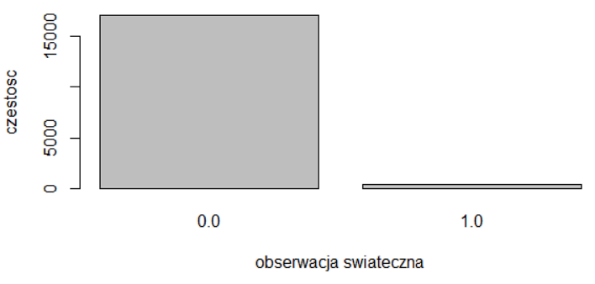
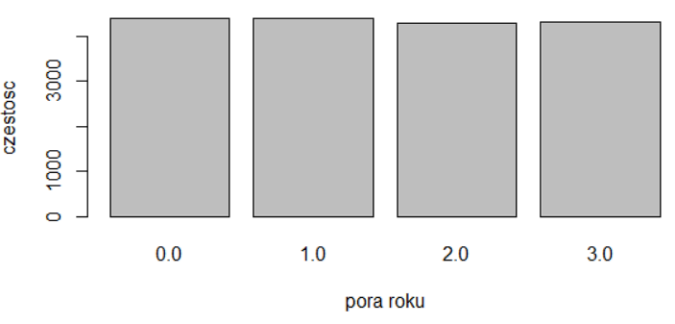
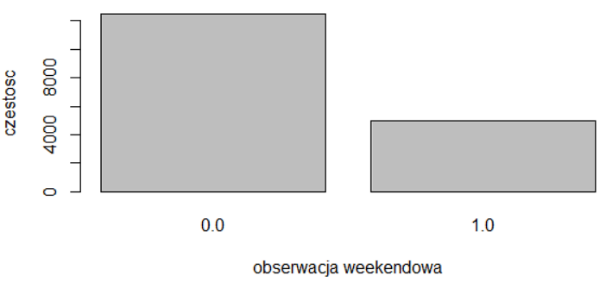
**Rysunek 25.** Histogram prędkości wiatru



Dodatkowo, zgodnie z obliczeniami wykonanymi przez R pierwszy kwartyl prędkości wiatru jest na poziomie 10 km/h, mediana to 15 km/h, zaś trzeci kwartyl to około 20,5 km/h.

Pogoda, wakacje, weekend, pory roku

W przypadku zmiennych kodowanych nie stwierdzamy żadnych uchybień. Wykresy częstości w przypadku tych zmiennych przedstawia Rysunek 26.

**Rysunek 26.** Wykresy zmiennych kodowanych

Komentarze do powyższych danych są następujące:

* Rodzaj pogody: deszcz z burzą i opady śniegu wystąpiły bardzo rzadko w rozważanym okresie, odpowiednio 14 i 60 razy, co stanowi 0,08% i 0,34% wszystkich obserwacji; najwięcej było bezchmurnych godzin – 6150 przypadków (35,32%).
* Obserwacje świąteczne stanowią zaledwie 2,2% wszystkich wskazań, dlatego zmienna dotycząca godzin świątecznych jest bardzo mało różnicująca i nie będzie brana pod uwagę w naszych modelach
* Weekendy powinny oczywiście być bardzo zbliżone lub równe 2/7 wszystkich obserwacji i tak rzeczywiście jest dla naszych danych (28,5%)
* Dane obejmują prawie dwuletni okres, więc pory roku rozkładają się niemal równomiernie.

Podsumowując, wykorzystywane w niniejszym projekcie dane są dobre oraz pełne. Występują jedynie jednostkowe błędy, obserwacje które są z nimi powiązane (nr 7027 i 15834) nie będą uwzględniane w dalszej analizie.

1. **Model**

Modelem rozważanym przez nas w tej części będzie Model 1., postaci:

,

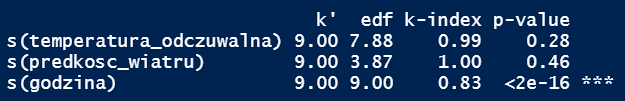
gdzie to pewne funkcje wygładzające.

Funkcje te (funkcja *s* w R) na etapie szacowania modelu wybierają optymalny stopień wygładzania na podstawie GCV (*Generalized Cross - Validation*).

Sprawdzimy na początku, czy parametr k jest dobrze dobrany w rozważanym modelu.

Niestety, wyniki uzyskane funkcją *gam.check,* przedstawione na Rysunku 27, nie napawają optymizmem.

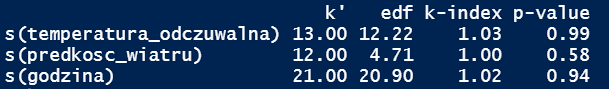
**Rysunek 27.** Wyniki funkcji *gam.check* dla Modelu 1.



Niskie p-value (wartości k<1) wskazuje na to, że k jest za małe (tym bardziej, że w przypadku godziny nie nastąpiła żadna redukcja wyjściowego wymiaru baz), w związku z czym należy dokonać modyfikacji w poszukiwaniu prawidłowej (wystarczająco dużej) wartości k.

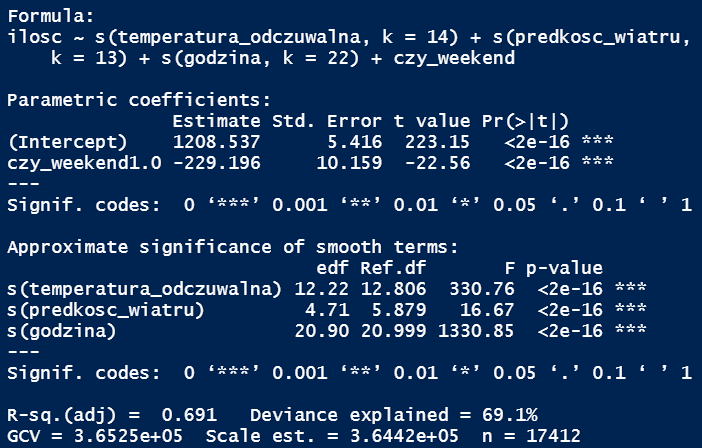
Dopiero przy równym kolejno 13, 12, 21 otrzymujemy rezultaty wskazujące na dobry wybór stopnia wygładzania (Rysunek 28.).

**Rysunek 28.** Wyniki funkcji *gam\_check* dla Modelu 1 ze zmodyfikowanym stopniem wygładzenia



Przejdziemy teraz do omówienia statystyk podsumowujących Model 1., które to zaprezentowane są na Rysunku 29.

**Rysunek 29.** Statystyki podsumowujące Modelu 1.

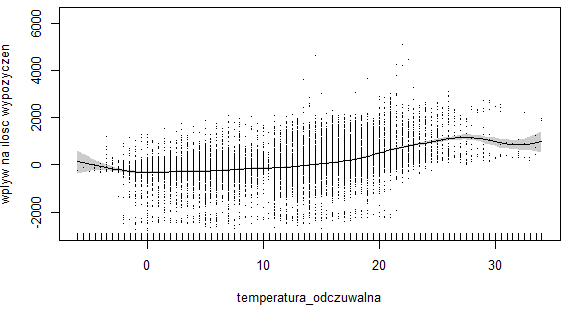


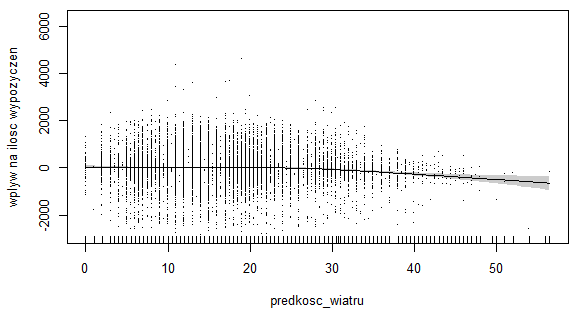
Wypiszemy najważniejsze spostrzeżenia, które możemy wysnuć na podstawie powyższego podsumowania:

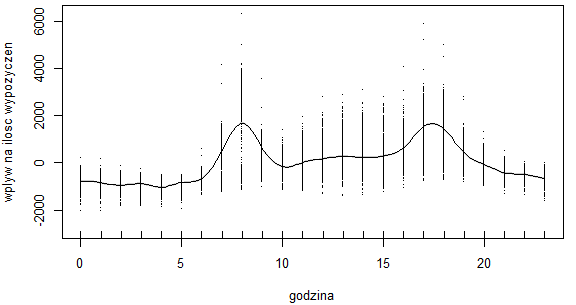
* Dla elementu parametrycznego w naszym równaniu, czyli zmiennej *czy\_weekend* (przypomnijmy, iż przyjmuje ona wartość 1, jeśli jest weekend i 0 w przeciwnym przypadku) możemy stwierdzić, że w weekendy liczba nowych wypożyczeń rowerów w ciągu godziny jest średnio o 229 mniejsza niż w pozostałe dni. Jest to zgodne z oczekiwaniami, gdyż w tygodniu ilość wypożyczanych rowerów powinna być większa ze względu na dojazd do pracy lub szkoły. Współczynnik ten jest istotny ze względu na wartość p bliską 0.
* Wszystkie składniki wygładzające (dlazmiennych *temperatura\_odczuwalna, predkosc\_wiatru* i *godzina*)istotnie wpływają na model ze względu na wartość p bliską 0 i wysoką wartość statystyki F; bardzo wysoka wartość tej statystyki dla zmiennej *godzina* świadczy o tym, że ta zmienna ma stosunkowo bardzo duży wpływ na ilość nowych wypożyczeń rowerów.
* Współczynnik jest na przywoitym poziomie 0,691. Oznacza to, że model jest dobrze dopasowany i wyjaśnia 69,1% zmienności liczby nowych wypożyczeń rowerów.
* GCV równe i skala na podobnym poziomie sugerują, że średni błąd predykcji takiego modelu wynosi około 603 wypożyczenia nowych rowerów w ciągu godziny. Porównując te liczbę z rzędami wielkości wypożyczeń rowerów, należy stwierdzić, że model nie nadaje się do celów predykcyjnych.

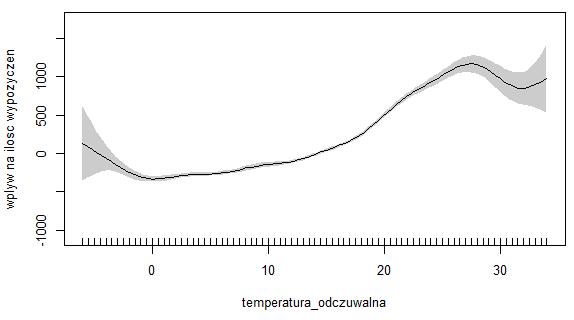
Następnym krokiem będzie ocena graficzna wielkości i kierunku wpływu każdej z wygładzonych zmiennych na podstawie wykresów reszt cząstkowych. Korzystając z funkcji *plot* w R, otrzymujemy wykresy przedstawione na Rysunku 30, gdzie pierwsze trzy wykresy to wykresy reszt cząstkowych, zaś kolejne to wykresy z wyłączonymi resztami, dzięki czemu będziemy mogli lepiej ocenić wpływ zmiennych niezależnych na modelowaną zmienną.

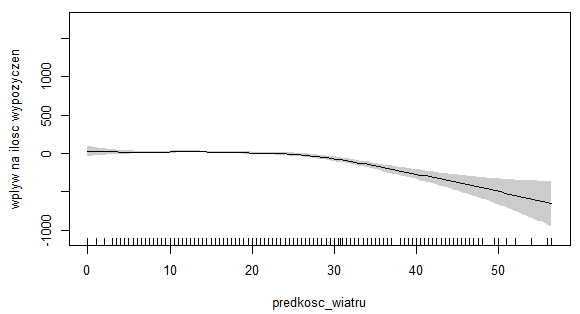
**Rysunek 30.** Wpływ poziomu zmiennych niezależnych na ilość nowych wypożyczeń rowerów w Modelu 1.

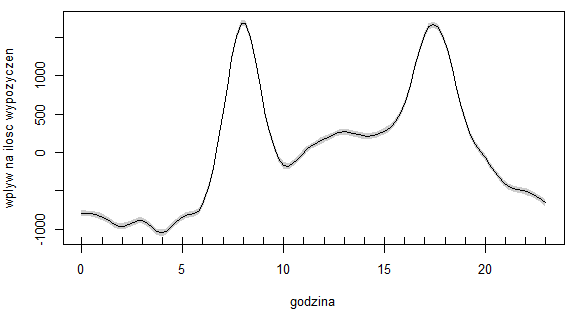


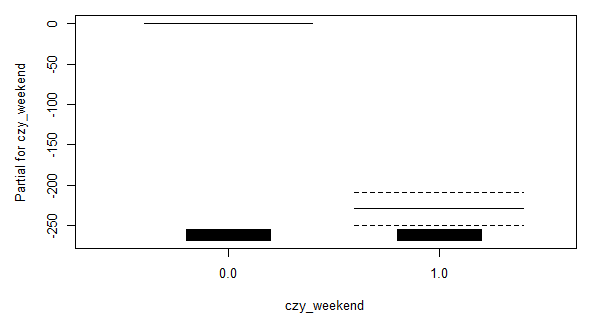












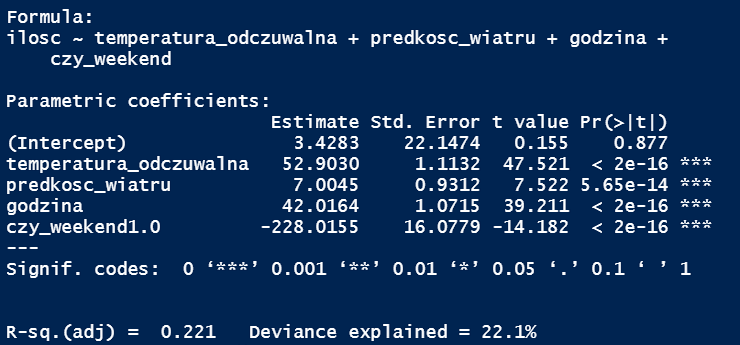
Na pierwszych trzech wykresach mamy przedstawione reszty cząstkowe, które zobrazowane są jako zbiory kropek nad i pod wygładzoną krzywą relacji. Widzimy miejsca, w których reszty te są dosyć mocno rozproszone, co świadczy to o tym, że Model 1. nie jest idealny (nasze zmienne nie wyjaśniają całej zmienności wypożyczeń rowerów), ale jest to model poprawny, ze względu na równomierne rozproszenie (nad i pod krzywą). Co więcej, widać również pewne miejsca (czyli piony), w których występują pojedyncze kropki lub ich zlepek przypomina kształtem odcinek o stosunkowo krótkiej długości, co z kolei sugeruje dobre dopasowanie (choć czasami może świadczyć o po prostu mniejszej w tym miejscu liczbie obserwacji). Na kolejnych trzech wykresach zobrazowane są oceny wpływu odpowiednio temperatury odczuwalnej, prędkości wiatru, czy też godziny na ilość nowych wypożyczeń. Szare pasy wokół wygładzonych przebiegów funkcji, wskazują na przedziały ufności. W tym miejscu dokonajmy interpretacji tych właśnie wykresów:

* Wraz ze wzrostem temperatury od 0 do około 27 stopni Celsjusza nieustannie rośnie również ilość nowych wypożyczeń rowerów. Najpierw, do około 18 stopni jest to wzrost łagodny, zbliżony do liniowego, zaś od 18 stopni zwiększa się jego tempo (w tych również okolicach mamy jeden z wyraźniejszych punktów przegięcia wykresu – funkcja z wypukłości przechodzi we wklęsłość). Spadki ilości wypożyczeń występują jedynie przy wzroście temperatury od ujemnej do 0 stopni oraz dla temperatur powyżej 27 stopni do około 31 stopni. Wizualnie, można stwierdzić, że najmniejszej ilości wypożyczeń towarzyszą temperatury w okolicach 0 stopni, zaś największej wysokie temperatury, około 27-28 stopni Celsjusza (mamy tutaj znaczący wzrost o ponad 1000 rowerów w stosunku do poziomu odniesienia). Wykres dosyć dobrze odzwierciedla ogólną zasadę, że im większa temperatura (ale nie upał), tym chętniej ludzie sięgają po rowery.
* Interpretacja wykresu obrazującego zależność prędkości wiatru od nowych wypożyczeń jest zgodna z oczekiwaniami. Dla wiatru do prędkości około 25 km/h ilość wypożyczeń jest niemal stała, taki wiatr nie różnicuje chęci ludzi do wsiadania na rowery. Następnie, dla wiatru o większej szybkości funkcja jest malejąca w sposób wypukły, równocześnie spadek ilości nowych wypożyczeń jest coraz większy.
* Wykres zależności godziny wypożyczenia rowerów od ilości wypożyczeń jest również intuicyjnie łatwo wytłumaczalny. Między godziną 22:00 a 6:00 ilość wypożyczeń jest na najmniejszym poziomie (ludzie wtedy siedzą w domach/śpią), jest wtedy minimalna populacyjna potrzeba korzystania z jakiejkolwiek komunikacji. Między godziną 6:00 a 8:30-9:00 następuje gwałtowny i nieustanny wzrost: o 8:30-9:00 ilość nowych wypożyczeń jest średnio o około 2000 sztuk większa niż o 6:00. Jest to oczywiście związane z porannym dojazdem ludzi do pracy i do szkoły. Podobne poziomy wypożyczeń osiągane są między godziną 17:00 a 18:00, gdy osoby te wracają do swoich domów.
* Ostatni wykres jest potwierdzeniem otrzymanego wcześniej współczynnika (Rysunek 29.) - w weekendy liczba nowych wypożyczeń rowerów w ciągu godziny jest średnio o 229 mniejsza niż w pozostałe dni. Na wykresie przerywaną linią poziomą oznaczono przedziały ufności.

Podsumowując, przedstawiony na wykresach wpływ prędkości wiatru, odczuwalnej temperatury, czy też godziny na ilość nowych wypożyczeń rowerów, jest zgodny z oczekiwaniami. Dodatkowo, Model 1. jest niezłym modelem ze względu na całkiem wysoki poziom wyjaśniania zmienności 69,1%.

Spróbujemy teraz porównać nasz model z modelem liniowym i formalnie ocenić, czy jest on modelem lepszym. Wyniki dla modelu liniowego przedstawiają się następująco (Rysunek 31.)

**Rysunek 31.** Podsumowanie modelu liniowego

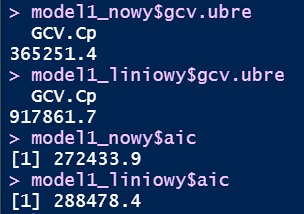


Przeprowadźmy krótką interpretację otrzymanych współczynników:

* Wzrost temperatury odczuwalnej o 1 stopień Celsjusza zwiększa średnio ilość nowych wypożyczeń rowerów (w przeciągu godziny) o 52,9 sztuki.
* Przy wzroście prędkości wiatru o 1km/h liczba wypożyczeń rowerów rośnie średnio o 7 sztuk.
* Współczynnik przy godzinie zostawiamy bez interpretacji (ze względu na cykliczność).
* W weekendy ilość wypożyczeń rowerów jest średnio o 228 sztuk mniejsza (na godzinę).

Model ten jest znacznie gorszy od Modelu 1. pod względem wyjaśniania zmienności liczby nowych wypożyczeń rowerów (22,1% vs 69,1%). Potwierdzenie wyższości Modelu 1. nad modelem liniowym otrzymujemy porównując miary dopasowania tych modeli (Rysunek 32.)

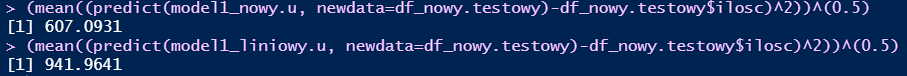
**Rysunek 32.** Porównanie miar dopasowania Modelu 1. (*model1\_nowy*) i modelu liniowego (*model1\_liniowy*)



Zarówno GCV jak i kryterium informacyjne Akaike są większe w przypadku modelu liniowego, zatem Model 1. jest modelem znacznie lepszym niż liniowy odpowiednik.

Na koniec dokonamy porównania dokładności predykcji Modelu 1. i modelu liniowego, dzieląc zbiór danych na próbę uczącą i testową. Po wygenerowaniu numerów testowych stanowiących ¼ wszystkich obserwacji (losowanie bez zwracania), otrzymujemy następujące wyniki średniego błędu dla Modelu 1. i modelu liniowego (Rysunek 33.):

**Rysunek 33.** Średni błąd predykcji Modelu 1. (górna wartość) i modelu liniowego



Potwierdzają się wcześniejsze obserwacje – model liniowy jest znacznie gorszy od Modelu 1. Średni błąd predykcji w przypadku modelu liniowego jest większy o 55,2% w porównaniu z Modelem 1.

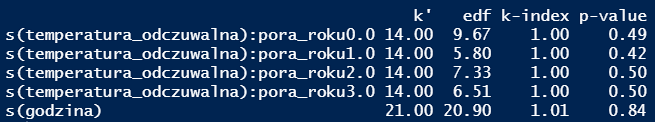
Pokrótce zaprezentujemy w tej części projektu jeszcze jeden model (Model 2.), który jest przykładem modelu z interakcją zmiennej wygładzanej ze zmienną czynnikową. Model 2. ma u nas następującą postać:

,

gdzie to pewne funkcje wygładzające.

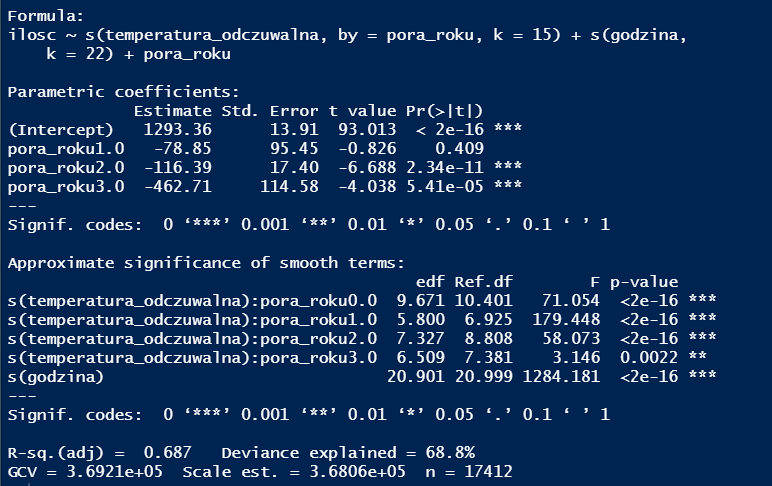
Istota interakcji polega w tym przypadku na tym, że chcemy modelować wpływ odczuwalnej temperatury na ilość nowych wypożyczeń rowerów w zależności od pory roku. Podobnie jak w Modelu 1. napotykamy tutaj na problemy w postaci zbyt niskiej wartości k, modyfikując jednak parametr k, dla którego k’ wynosi odpowiednio 14 i 21 otrzymujemy wystarczający stopień wygładzania (Rysunek 34.).

**Rysunek 34.**  Wyniki funkcji *gam.check* dla Modelu 2. ze zmodyfikowanym stopniem wygładzania



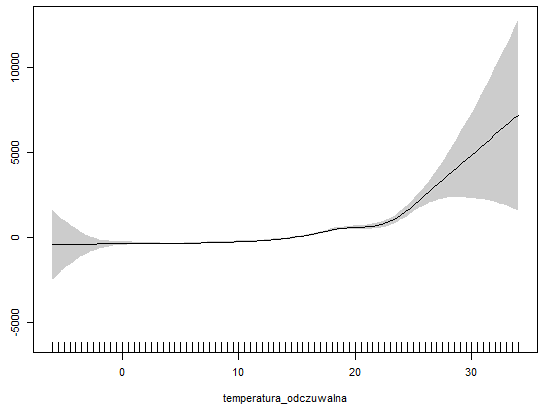
Mamy bowiem k-index nie mniejszy niż 1 oraz wysoki wartości p-value. Statystyki podsumowujące Model 2. Przedstawiamy na kolejnym rysunku (Rysunek 35.)

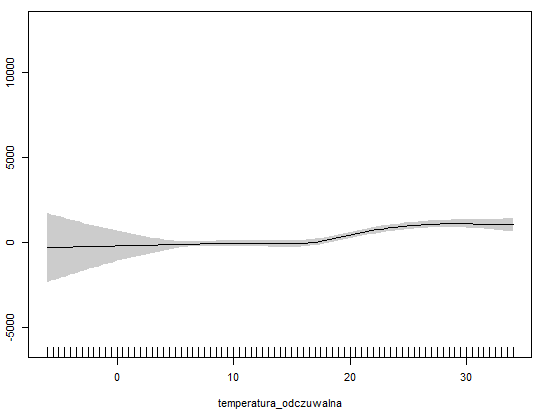
**Rysunek 35.** Statystyki podsumowujące dla Modelu 2.

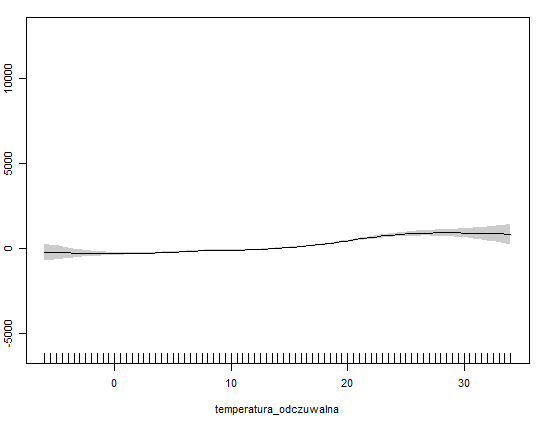


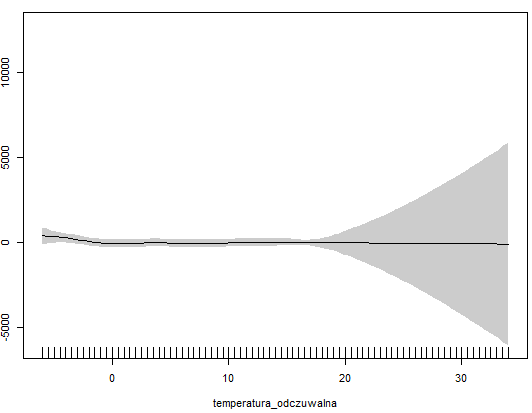
Zauważmy przede wszystkim, że otrzymane składniki są istotne (małe p-value, wysokie wartość statystyki F), z kolei parametr przy *pora\_roku1.0* nie jest istotny. Co więcej, model wyjaśnia 68,8% zmienności zmiennej objaśnianej, co jest dobrym rezultatem (na bardzo zbliżonym poziomie, co w przypadku Modelu 1.). Widzimy również, że latem, jesienią i zimą średnia liczba wypożyczeń rowerów na godzinę jest mniejsza od wypożyczeń na wiosnę o odpowiednio około 79, 116 i 463 sztuki. Jest to rozsądny wynik – zimą rowery mają mniejszą liczę użytkowników, a wiosna wydaje się być optymalną porą roku do jazdy rowerem. Zerknijmy jeszcze na najważniejsze wykresy, czyli te związane z wpływem odczuwalnej temperatury na ilość wypożyczeń nowych rowerów w zależności od pory roku (Rysunek 36.)

**Rysunek 36.** Wpływ odczuwalnej temperatury na ilość wypożyczeń, w zależności od pory roku (kolejno: wiosna, lato, jesień, zima)









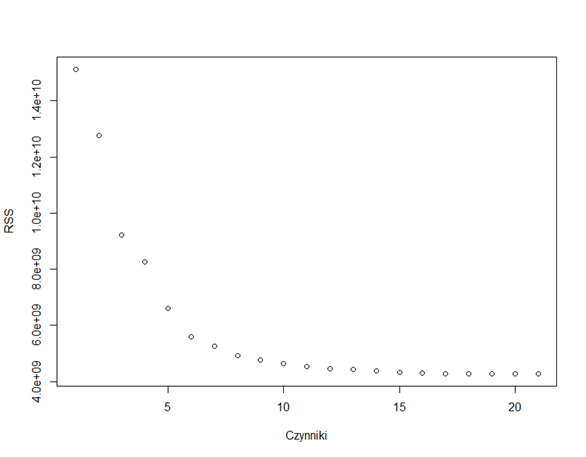
Na podstawie powyższych wykresów widzimy, że wspomniany wpływ bywa różnicowany w zależności od pory roku. O ile trzy ostatnie wykresy mają zbliżony przebieg – temperatura odczuwalna ma niewielki wpływ na ilość wypożyczeń latem, jesienią i zimą, o tyle przebieg funkcji związanej z wiosną jest inny niż pozostałe. Wprawdzie dla temperatur do około 20 stopni nie widać wielkich różnic w ilości wypożyczonych rowerów (jest delikatny wzrost), o tyle powyżej tej temperatury wzrost ilości wypożyczeń wiosną może gwałtownie rosnąć (mamy tutaj bardzo szeroki przedział ufności). Ten prawdopodobny wzrost byłby czymś naturalnym. Dobra wakacyjna pogoda wydaje się być najlepszym stanem natury na, okazjonalne lub nie, przejażdżki rowerowe. Wiosną wiele osób budzi się do życia po zimowych krótkich dniach, a słońce (częste wiosną) i ciepło (ale nie upał) powodują, że ludzie czują przypływ energii, który z radością wykorzystują poprzez uprawianie sportu, w tym również przejażdżki rowerem.

**Część 3. Model MARS dla liczby wypożyczeń rowerów w Londynie**

W części trzeciej estymowane będą modele MARS, które podobnie jak modele GAM, korzystają z przekształceń zmiennych egzogenicznych za pomocą funkcji bazowych, w taki sposób, aby modelować nieliniowe zjawisko liniowymi metodami regresji. Dane użyte w kolejnych modelach, to te same dane, który były użyte w części drugiej.

Pierwszy estymowany model MARS to model addytywny bez interakcji. Przy domyślnym ustawieniu parametru maksymalnej liczby czynników w modelu (funkcji bazowych zmiennych egzogenicznych oraz wyrazu wolnego), początkowo zostało dobranych 21 czynników. Następnie wykluczono z modelu dwie najmniej istotne funkcje bazowe, które w najmniejszym stopniu zwiększają sumę kwadratów reszt, a model z pozostałymi 19 czynnikami charakteryzuje się mniejszą wartością GCV. Wykluczone funkcje bazowe to h(23 – prędkość wiatru) oraz h(godzina – 11), a więc czynniki mówiące o wpływie wzrostu prędkości wiatru gdy ten jest niższy od 23 km/h, oraz o wpływie upływu kolejnych godzin po godzinie 11.

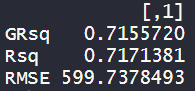
**Rysunek 37.** Suma kwadratów reszt dla danej ilości czynników w modelu



Wśród 19 czynników modelu MARS1 wykorzystane zostały jedynie 6 zmiennych objaśniających. Są to zmienne: pogoda, czy\_weekend, temperatura\_odczuwalna, wilgotność, prędkość wiatru oraz godzina. A więc zmienne temperatura, czy\_swieto i pora\_roku zostały uznane za nieistotne i nie są ujęte w modelu.

W poniższej tabeli znajdują się miary GRsq – miara, która przybiera wartości tym wyższe im mniejsze jest GCV modelu względem modelu tylko ze stałą, Rsq – współczynnik determinacji, mówiący o tym, jak duża część zmienności w danych wyjaśniona jest zmiennością zmiennych niezależnych, oraz RMSE – średni błąd kwadratowy na próbie uczącej (tej samej co w części drugiej). Te miary posłużą do porównania kolejnych modeli MARS pomiędzy sobą oraz z modelem GAM (model1) – którego wartość RMSE na próbie uczącej wyniosła 607,09.

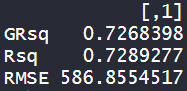
**Rysunek 38.** Miary oceny modelu MARS1



Model na podstawie RMSE zdaje się być podobnie efektywny w celach prognozy co model GAM (model1), a jego Rsq jest nieco wyższe. Dokładne interpretacje parametrów modelu zostaną opisane jedynie dla tego o najlepszych miarach ocen dopasowania i predykcji.

Następnie rozważony został model z interakcjami, w tym celu do modelu MARS2 włączone zostały interakcje do trzeciego stopnia. Model początkowo został estymowany z 23 czynnikami, a następnie zmniejszył ich liczbę do 21. Zmienne użyte w modelu są te same co w MARS1, z wyjątkiem zmiennej prędkość wiatru, która nie została wykorzystana.

**Rysunek 39.** Miary oceny modelu MARS2

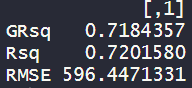


Dodanie interakcji do 3 stopnia w modelu spowodowało wzrost liczby czynników w modelu MARS2, a także wzrost GRsq oraz Rsq o około 1 punkt procentowy, a błąd prognozy RMSE na próbie testowej zmniejszył się o około 13 rowerów. Najczęściej występujące interakcje w tym modelu to te pomiędzy godziną oraz wilgotnością i temperaturą odczuwalną.

Kolejnym rozważonym modelem jest MARS3, model addytywny ze zmienioną liczbą parametrów minspan – minimalnej przerwy pomiędzy wiązaniami funkcji bazowych oraz maksymalnej liczby czynników przed ich przycinaniem. Metoda symulacji wykazała, że dla tej próby uczącej optymalną wartością maksymalnej ilości czynników jest 37 (z zakresu 20 do 50), a dobór wartości parametru minspan w przedziale 1 do 20 nie ma istotnego wpływu na model. Wybrane parametry minimalizują GCV modelu.

W modelu MARS3, podobnie jak w poprzednim modelu, uwzględnione zostało 21 czynników, jest liczba wykorzystanych zmiennych egzogenicznych jest większa – zostały wykorzystane wszystkie zmienne z modelu MARS1 oraz zmienna czy\_swieto i pora roku, a więc wszystkie zmienne poza zmienną temperatura.

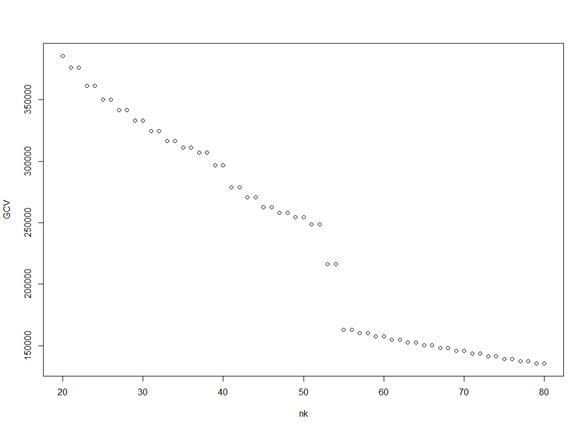
**Rysunek 40.** Miary oceny modelu MARS3



Zmiany tych parametrów spowodowały wzrost liczby czynników, wykorzystanych zmiennych, a także lekki wzrost miar GRsq, Rsq oraz spadek RMSE. Jednak wszystkie 3 miary oceny są wciąż gorsze, od tych w modelu z interakcjami.

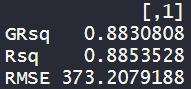
Do estymacji ostatniego modelu MARS4, użyta została ta sama procedura optymalizacji parametrów, ale dla modelu z interakcjami, oraz z dodatkowym parametrem ustalonym metodą symulacji jest poziom interakcji. Najpierw ustalony został maksymalny poziom interakcji – z zakresu od 1 do 10, najniższą wartość GCV wykazał modelu o maksymalnie 3 poziomie interakcji – kolejne zwiększenie poziomu maksymalnej interakcji nie zmniejszają błędu GCV. Minimalne GCV ze względu na maksymalną liczbę czynników przed przycinaniem w zakresie od 20 do 80, osiągnął model z wartością parametru nk równą 79. A wartość minspan ponownie nieznacznie wpływa na wartości GCV modelu, jednak tą optymalną w tym przypadku jest 3.

**Rysunek 41.** Wpływ parametru nk na GCV modelu MARS4



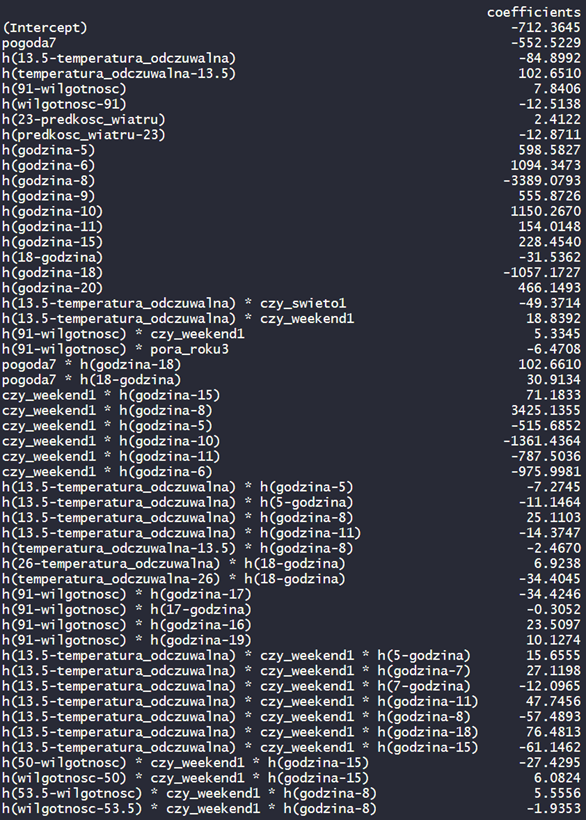
Model MARS4 pomimo długiej estymacji związanej z szukaniem optymalnych parametrów, charakteryzuje się najlepszymi wartościami GRsq, Rsq oraz RMSE na próbie testowej ze wszystkich modeli (w tym modeli GAM).

**Rysunek 42.** Miary oceny modelu MARS4



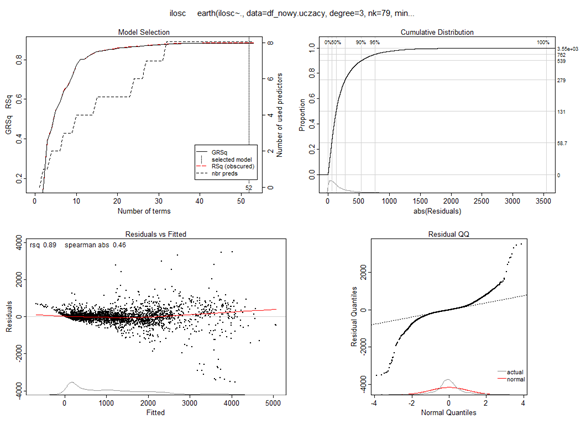
Model wykorzystuje wszystkie zmienne objaśniające poza zmienną ‘temperatura’, oraz uwzględnia aż 52 czynniki (funkcje bazowe z interakcjami lub bez, oraz wyraz wolny). Większość funkcji bazowych dotyczy godziny wypożyczenia rowerów, kombinacji godziny i tego czy jest weekend, oraz kombinacji godziny i wilgotność czy też temperatury odczuwalnej. Na rysunku 43. Wypisane zostały wszystkie parametry modelu, które ze względu na ich duża ilość i interakcje są trudne do interpretacji.

**Rysunek 43.** Parametry strukturalne modelu MARS4



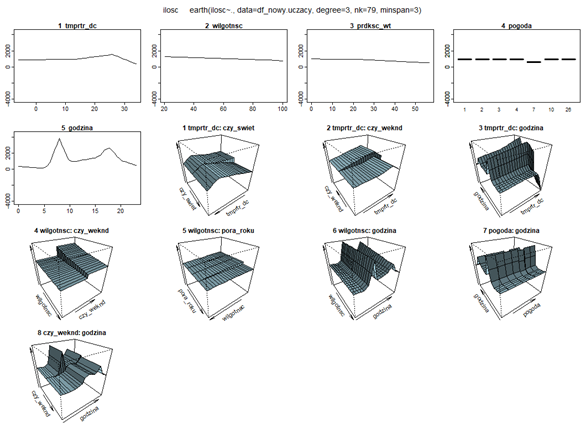
Upływ czasu o godzinę, po godzinie 8, zmniejsza ilość wypożyczonych rowerów średnio o około 3389 rowerów, przy założeniu stałości pozostałych czynników, ale ze względu na występowania w modelu wielu funkcji bazowych związanych z godziną, ta interpretacja jest mało trafna, ponieważ inne czynniki także muszą mieć wpływ – wypożyczenia rowerów wzrastają po godzinie 5 średnio o około 598 co godzinę oraz kolejne 1094 co każdą godzinę po godzinie 6. Wiele czynników uwzględnia także interakcje, upływ 1 godziny po 8 rano w weekend zwiększa liczbę wypożyczonych rowerów średnio około 3425 rowerów przy braku wpływu pozostałych zmiennych.

**Rysunek 44.** Rozkład reszt modelu MARS4



Model stosunkowo dobrze prognozuje generalne zachowanie się zmiennej objaśnianej, największe błędy dotyczą obserwacji odstających, klasyfikacja których może być związana z elementem stochastycznym, lub nieuwzględnioną w badaniu zmienną objaśniającą. Ogólne dopasowanie modelu do danych jest stosunkowo dobre. Błąd RMSE wynoszący około 373 rowery przy średniej ilości wypożyczonych rowerów w ciągu danej godziny danego dnia około 1163 rowery daje dosyć duży błąd względem średniej, ale jest to związane z bardzo dużą zmiennością zmiennej objaśnianej.

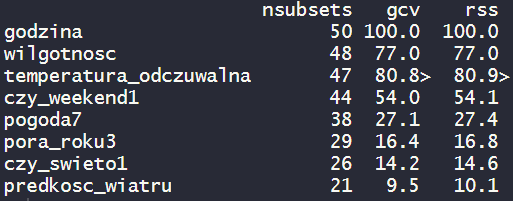
**Rysunek 45.** Wpływ poszczególnych zmiennych objaśniających na zmienną objaśnianą



Ze względu na interakcje wpływ poszczególnych zmiennych jest trudny w interpretacji, oceniając zmienne bez interpretacji, można zauważyć, że ilość wypożyczonych rowerów wzrasta od poranku do południa oraz wzrasta wieczorem, przed godziną 20. Wilgotność oraz prędkość wiatru wpływają na zmniejszenie ilości wypożyczanych rowerów, a temperatura odczuwalna wpływa na wzrost tej ilości (szczególnie powyżej temperatury 13,5 stopni Celsjusza), aż do temperatury 26 stopni Celsjusza, gdy kolejny wzrost temperatury zmniejsza prognozowaną ilość wypożyczeni. Wartość zmiennej pogoda 7, która oznacza deszcz, zmniejsza prognozowaną liczbę wypożyczeni średnio o około 553 rowery. Zmienne weekend oraz święto dodatkowo wpływają na wzrost lub spadek wypożyczeń rowerów przy odpowiedniej temperaturze odczuwalnej i godzinie – przy dużej temperaturze w godzina wieczornych modelowany jest wzrost, podczas gdy w godzinach porannych powoduje spadek.

Na rysunku 46. Przedstawiona jest ważność zmiennych objaśniających w modelu ze względu na ilość wystąpień zmiennych w modelach ze wcześniej fazy estymacji, wpływ na zmniejszenie wartości błędów GCV i RSS.

**Rysunek 46.** Ważność poszczególnych predyktorów w modelu MARS4



Najbardziej istotnymi zmiennymi dla modelu jest godzina, wilgotność, temperatura odczuwalna oraz to czy jest weekend.