2 czerwca 2021

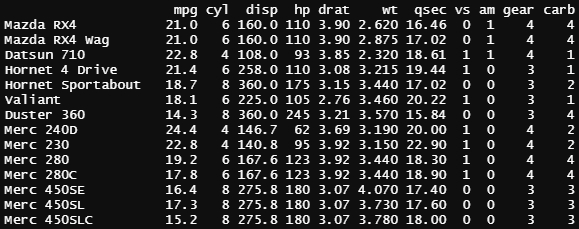
Kinga Kalisz & Marek Pałdyna

Porównanie Metod Regresji

Pakiet R

# MTCars

## Opis danych



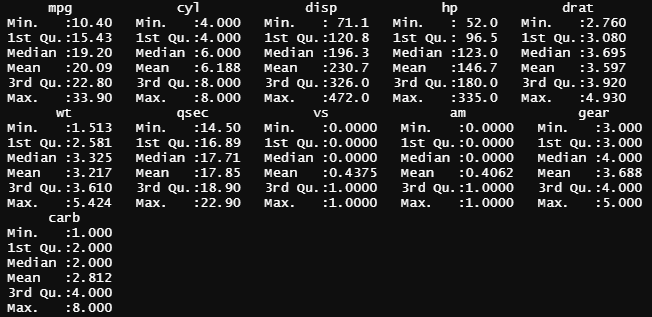
Zbiór danych MTCars zawarty jest w zbiorach danych środowiska R.   
Składa się on z 11 kolumn i przedstawia dane na temat 32 samochodów z lat 1973-74. Zaciągnięty został z „Motor Trend US magazine”.

Za cel przewidywań na tym zbiorze została obrana kolumna „qsec”, która opisuje czas na ćwierć mili.

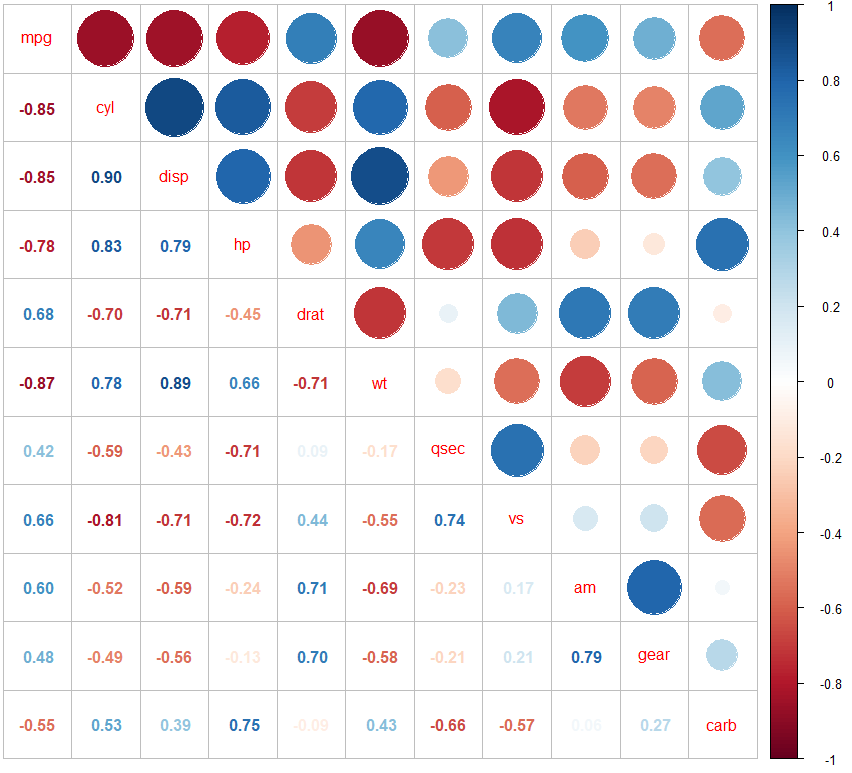
Kolumny:

* „mpg” – spalanie paliwa (mil/galon)
* „cyl” – liczba cylindrów
* „disp” – pojemność silnika (cu.in.)
* „hp” – moc silnika (KM)
* „drat” – przełożenie tylnej osi
* „wt” – waga (1000 lbs)
* „qsec” – czas na ¼ mili
* „vs” – typ silnika (0 – w kształcie V, 1 - prosty)
* „am” – skrzynia biegów (0 – automatyczna, 1 -manualna)
* „gear” – liczba biegów do jazdy do przodu
* „carb” – liczba gaźników

## Podsumowanie danych:



## Korelacja zmiennych:



W tym zbiorze danych istnieją duże korelacje „pozytywne” i „negatywne”.

Korelacje negatywne zachodzą pomiędzy:

* spalaniem i ilością cylindrów silnika,
* spalaniem i wagą samochodu
* spalaniem i pojemnością silnika
* liczbą cylindrów i typem silnika

Korelacje pozytywne zachodzą pomiędzy:

* pojemnością silnika, a wagą samochodu
* pojemnością silnika, a liczbą cylindrów
* rodzajem silnika a liczbą biegów do jazdy do przodu
* liczbą cylindrów, a mocą silnika

## Obróbka danych

Do najlepszego dopasowania danych do zadanego problemu użyto funkcji step. Dobrała ona kolumny tak aby regresja wychodziła jak najlepsza -> najlepsze R-squared. Funkcja ta ma 3 „metody” – forward, backward i both.

Wartości squared dla danych:

* forward r.squared = 0.8523863
* backward r.squared = 0.8453189
* **both r.squared = 0.8642928**

Wybrany został więc model “both” –> „qsec ~ cyl + disp + wt + vs + am + carb”

## Modele regresji

Zostały zaimplementowane 3 modele regresji:

* LM
* GLM
  + Gaussian
  + Poisson
  + Gamma
  + Quasi
  + Quasipoisson
* SVM z kernelami: (z biblioteki: e1071)
  + Linear
  + Polynomial
  + Radial
  + Sigmoid

## Porównanie poszczególnych modeli i ich błędów RMSE

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Typ | RMSE | MSE |
| LM | - | 3.073302 | 9.445186 |
| SVM | Linear | 0.9533641 | 0.9089031 |
| Polynomial | 1.548521 | 2.397916 |
| Radial | 1.575306 | 2.48159 |
| Sigmoid | 1.352628 | 1.829604 |
| GLM | Gaussian | 3.073302 | 9.445186 |
| Gamma | 17.91287 | 320.8708 |
| Quasipoisson | 15.02715 | 225.8152 |
| Quasi | 3.073302 | 9.445186 |
| Poisson | 15.02715 | 225.8152 |

## Podsumowanie

Najlepszym modelem dla danych MtCars jest model SVM Linear (funkcja kernelowa liniowa), osiągnął on najniższy poziom błędu RMSE = 0,7227685.   
Niewiele gorszy wynik dla powyższych danych osiągnął model SVM Radial (funkcja kernelowa radialna), osiągnął on poziom błędu RMSE = 0,7340955.

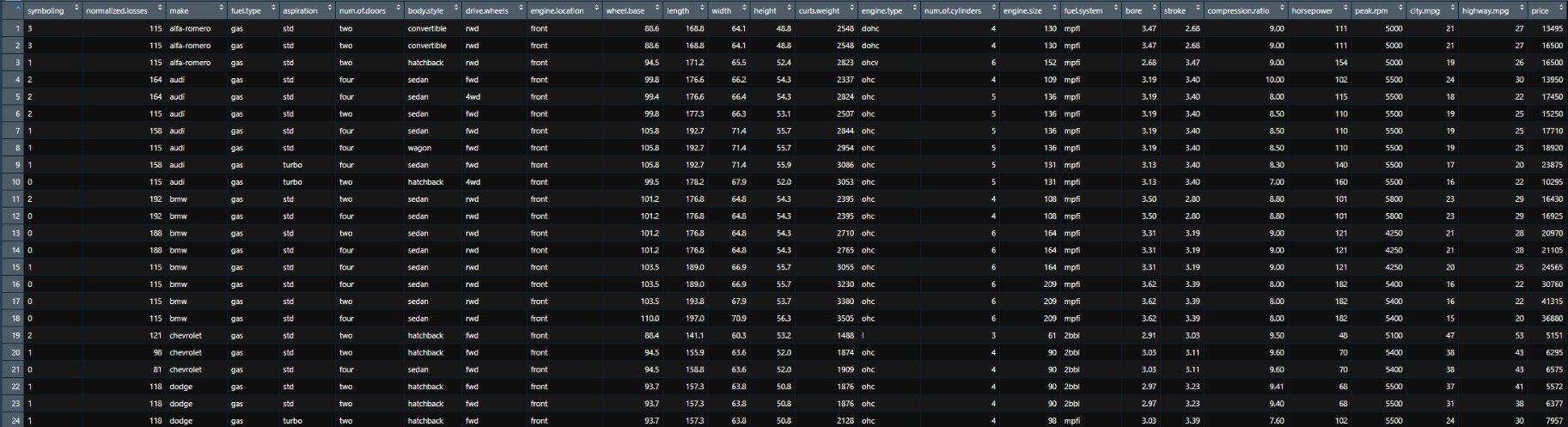
Najgorszym modelem dla danych MtCars jest model GLM Gamma, osiągnął on najwyższy poziom błędu RMSE = 17,86288.

# Imports-85

## Dane

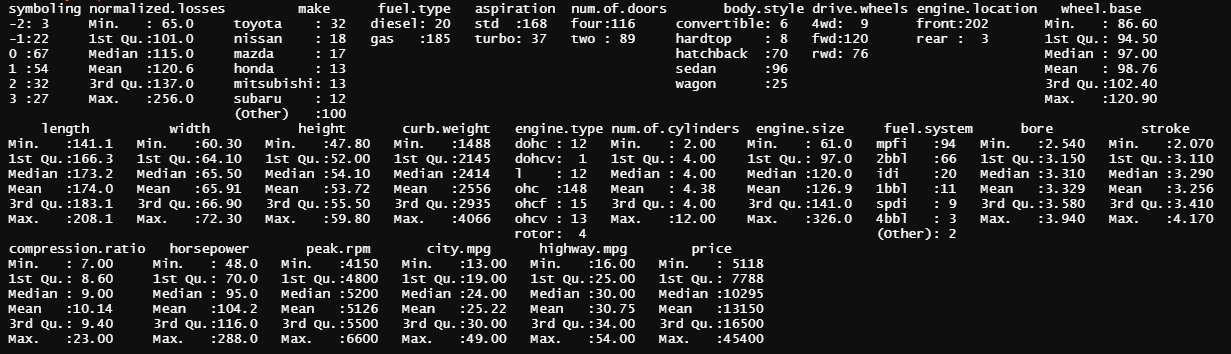
Plik imports-85.csv to plik zawierający dane o samochodach. Przechowuje 205 obserwacji. Dane pochodzą z zasobów udostępnionych przez dr inż. Kołopotka.

Kolumny:

* symbolling – poziom ryzyka samochodu
* normalized-losses – względna średnia wysokość odszkodowania za rok ubezpieczenia pojazdu
* make – marka
* fuel-type – typ paliwa
* aspiration – rodzaj ssania silnika
* num-of-doors – liczba drzwi
* body-style - typ nadwozia
* drive-wheels - rodzaj napędu
* engine-location - położenie silnika
* wheel-base - rozstaw osi
* length - dłogość
* width - szerokość
* height - wysokość
* curb-weight – masa własna
* engine-type – typ silnika
* num-of-cylinders – ilość cylindrów
* engine-size: rodzaj silnika
* fuel-system: system paliwowy
* bore: prześwit
* stroke: suw silnika
* compression-ratio: stopień sprężenia
* horsepower: moc silnika w KM
* peak-rpm: maksymalna ilość obrotów/min silnika
* city-mpg: zużycie benzyny w mieście
* highway-mpg: zużycie benzyny na drodze szybkiego ruchu
* price - cena

Na podstawie danych będziemy przeprowadzać przewidywanie „poziomu ryzyka ubezpieczenia samochodu” (symbolling) na podstawie innych parametrów.

## Podsumowanie danych



## Korelacja zmiennych

W tym zbiorze danych istnieją duże korelacje „pozytywne” i „negatywne”.

Korelacje negatywne istnieją pomiędzy:

* spalaniem w mieście a ilością KM
* spalaniem w mieście/na szosie a wagą własną
* współczynnikiem kompresji a rodzajem paliwa
* spalaniem a rodzajem paliwa

Korelacje pozytywne istnieją pomiędzy:

* rozstawem osi a długością
* długością a szerokością
* rozmiarem silnika a masą własną
* zużyciem na szosie a zużyciem w mieście

## Obróbka danych

Do najlepszego dopasowania danych do zadanego problemu użyto funkcji step. Dobrała ona kolumny tak aby regresja wychodziła jak najlepsza -> najlepsze R-squared. Funkcja ta ma 3 „metody” – forward, backward i both.

Wartości squared dla danych:

* forward r.squared = 0.7512137
* **backward r.squared = 0.8202633**
* both r.squared = 0.7014353

Wybrany został więc model “backward” –> „symboling ~ normalized.losses + make + num.of.doors + drive.wheels + length + width + engine.type + num.of.cylinders + engine.size + fuel.system + bore + city.mpg + highway.mpg”

## Modele regresji

Zostały zaimplementowane 3 modele regresji:

* LM
* GLM
  + Gaussian
  + Poisson
  + Gamma
  + Quasi
  + Quasipoisson
* SVM z kernelami:
  + Linear
  + Polynomial
  + Radial
  + Sigmoid

## Porównanie poszczególnych modeli i ich błędów RMSE

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Typ | RMSE | MSE |
| LM | - | 1.107603 | 1.226783 |
| SVM | Linear | 0.9823258 | 0.964964 |
| Polynomial | 1.079676 | 1.1657 |
| Radial | 0.9178603 | 0.8424676 |
| Sigmoid | 1.024627 | 1.04986 |
| GLM | Gaussian | 1.107603 | 1.226783 |
| Gamma | 3.776119 | 14.25907 |
| Quasipoisson | 2.827806 | 7.996488 |
| Quasi | 1.107603 | 1.226783 |
| Poisson | 2.827806 | 7.996488 |

## Podsumowanie

Najlepszym modelem dla danych z pliku imports-85.csv jest model SVM Radial (funkcja kernelowa radialna), osiągnął on najniższy poziom błędu RMSE = 0.5222752.

Najgorszym modelem dla danych z pliku imports-85.csv jest model GLM Gamma, osiągnął on najwyższy poziom błędu RMSE = 3.776119.

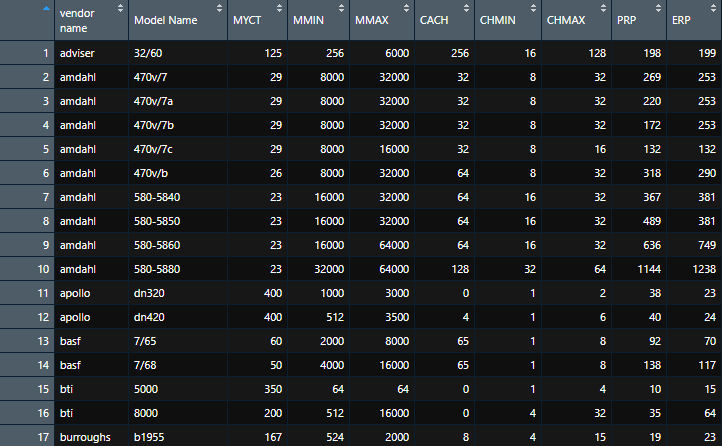
# Machine

## Dane

Plik machine.data zawiera dane na temat Komputerów PC. Posiada 209 wierszy danych. Dane zostały zaciągnięte z bazy: „<https://archive.ics.uci.edu/ml/datasets/Computer+Hardware>”

Kolumny:

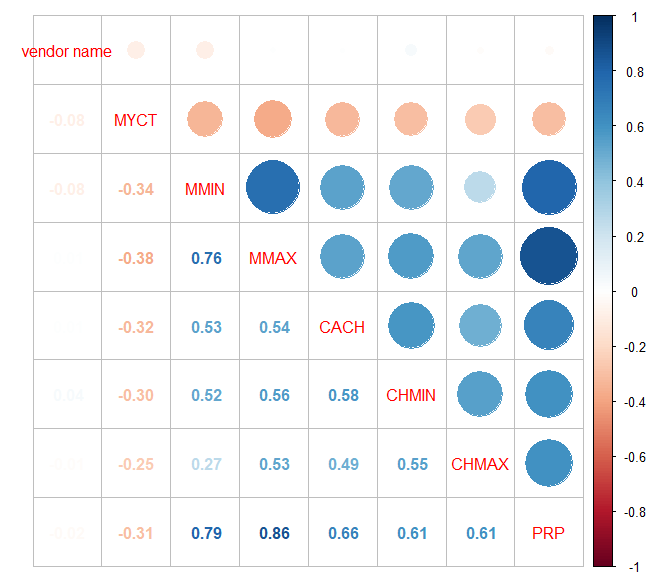
1. vendor name: nazwa dostawcy
2. Model Name: nazwa modelu
3. MYCT: czas cyklu maszyny w nanosekundach
4. MMIN: minimalna główna pamięć w kilobajtach
5. MMAX: maksymalna główna pamięć w kilobajtach
6. CACH: pamięć cache w kilobajtach
7. CHMIN: minimalna liczba kanałów w jednostkach
8. CHMAX: maksymalna liczba kanałów w jednostkach
9. PRP: opublikowana wydajność względna
10. ERP: przewidziana wydajność względna z oryginalnego artykułu



Na podstawie danych będziemy przeprowadzać przewidywanie „mocy względna” (PRP) komputera.

## Podsumowanie danych

## Korelacja zmiennych



W tym zbiorze danych istnieją duże korelacje „pozytywne” i słabe „negatywne”.

Korelacje pozytywne istnieją pomiędzy:

* maksymalną/minimalną główną pamięcią, a opublikowana wydajność względna
* minimalną, a maksymalną główną pamięcią

Korelacje negatywne istnieją pomiędzy:

* czasem cyklu maszyny, a maksymalną/minimalną główną pamięcią
* czasem cyklu maszyny, a pamięcią cache

Zmienna vendor name jest bardzo słabo skorelowana. Dodatkowo z danych zostały usunięte dwie zmienne:

* Model name – usunięta ze względu, że jest to zmienna unikalna dla każdego rekordu w danych
* ERP – wartość przewidywana i bardzo podobna do ERP więc będzie zaburzać pracę algorytmów

## Obróbka danych

Do najlepszego dopasowania danych do zadanego problemu użyto funkcji step. Dobrała ona kolumny tak aby regresja wychodziła jak najlepsza -> najlepsze R-squared. Funkcja ta ma 3 „metody” – forward, backward i both.

Wartości squared dla danych:

* forward r.squared = 0.8648239
* backward r.squared = 0.8648239
* both r.squared = 0.8648239

Dla tych danych wszystkie metody doboru zmiennych wybrały ten sam model -> „PRP ~ MYCT + MMIN + MMAX + CACH + CHMAX”

## Modele regresji

Zostały zaimplementowane 3 modele regresji:

* LM
* GLM
  + Gaussian
  + Poisson
  + Gamma
  + Quasi
  + Quasipoisson
* SVM z kernelami:
  + Linear
  + Polynomial
  + Radial
  + Sigmoid

## Porównanie poszczególnych modeli i ich błędów RMSE

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Typ | RMSE | MSE |
| LM | - | 79.34621 | 6295.821 |
| SVM | Linear | 78.81665 | 6212.064 |
| Polynomial | 780.6909 | 609478.3 |
| Radial | 145.6628 | 21217.65 |
| Sigmoid | 220.1345 | 48459.22 |
| GLM | Gaussian | 79.34621 | 6295.821 |
| Gamma | 205.0154 | 42031.3 |
| Quasipoisson | 202.0809 | 40836.68 |
| Quasi | 79.34621 | 6295.821 |
| Poisson | 202.0809 | 40836.68 |

## Podsumowanie

Najlepszym modelem dla danych z pliku machine.data jest model SVM Polynomial (funkcja kernelowa wielomianowa), osiągnął on najniższy poziom błędu RMSE = 39.50634.

Najgorszym modelem dla danych z pliku machine.data jest model SVM Sigmoid (funkcja kernelowa sigmoidalna), osiągnął on najwyższy poziom błędu RMSE = 525.4628

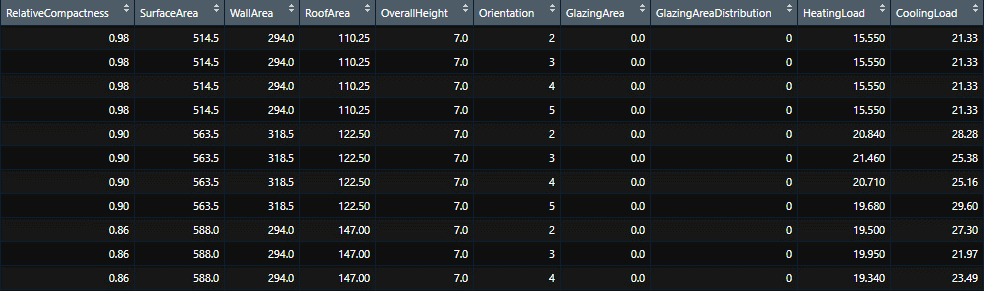
# Energy efficiency

## Dane

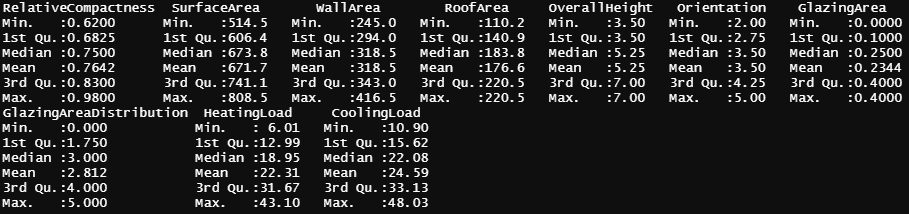
Plik ENB2012\_data.xlsx zawiera dane na temat Komputerów PC. Posiada 768 wierszy danych. Dane zostały zaciągnięte z bazy: „<https://archive.ics.uci.edu/ml/datasets/Energy+efficiency>”

Kolumny:

1. Relative Compactness – względna zwartość
2. Surface Area – powierzchnia budynku
3. Wall Area – powierzchnia ścian
4. Roof Area – powierzchnia dachu
5. Overall Height – całkowita wysokość
6. Orientation - orientacja
7. Glazing Area - przeszklenie
8. Glazing Area Distribution – rozmieszczenie powierzchni szkła
9. Heating Load – obciążenie ogrzewania
10. Cooling Load – obciązenie chłodzenia

Kolumny 1-8 to atrybuty budynków; 9-10 to dane przewidywane / do przewidywania na tym zbiorze danych. Wybraliśmy przewidywanie „Heating Load”

## Podsumowanie danych



## Korelacja zmiennych



W tym zbiorze danych istnieją duże korelacje „pozytywne” i duże „negatywne”.

Korelacje pozytywne istnieją pomiędzy:

* wysokością całkowitą, a obciążeniem ogrzewania
* wysokością całkowitą, a obciążeniem chłodzenia
* powierzchnią dachu, a powierzchnią budynku

Korelacje negatywne istnieją pomiędzy:

* powierzchnią dachu, a wysokością całkowitą
* względną zwartością, a zajmowaną powierzchnią
* względną zwartością, a powierzchnią dachu

Zmienna orientation jest nie skorelowana.

## Obróbka danych

Z danych została usunięta zmienna „Cooling Load”, ponieważ jest to zmienna „przewidziana” z dostępnych danych i nie powinna być użyta do przewidywania naszej zmiennej „Heating Load”.

Do najlepszego dopasowania danych do zadanego problemu użyto funkcji step. Dobrała ona kolumny tak aby regresja wychodziła jak najlepsza -> najlepsze R-squared. Funkcja ta ma 3 „metody” – forward, backward i both.

Wartości squared dla danych:

* forward r.squared = 0.9161955
* backward r.squared = 0.9161955
* both r.squared = 0.9161955

Dla tych danych wszystkie metody doboru zmiennych wybrały ten sam model: „HeatingLoad ~ RelativeCompactness + SurfaceArea + WallArea + OverallHeight + GlazingArea + GlazingAreaDistribution”

## Modele regresji

Zostały zaimplementowane 3 modele regresji:

* LM
* GLM
  + Gaussian
  + Poisson
  + Gamma
  + Quasi
  + Quasipoisson
* SVM z kernelami:
  + Linear
  + Polynomial
  + Radial
  + Sigmoid

## Porównanie poszczególnych modeli i ich błędów RMSE

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Typ | RMSE | MSE |
| LM | - | 3.012848 | 9.077253 |
| SVM | Linear | 3.098471 | 9.60052 |
| Polynomial | 3.657129 | 13.37459 |
| Radial | 2.548747 | 6.496113 |
| Sigmoid | 23.71798 | 562.5424 |
| GLM | Gaussian | 3.012848 | 9.077253 |
| Gamma | 24.88738 | 619.3818 |
| Quasipoisson | 22.00171 | 484.075 |
| Quasi | 3.012848 | 9.077253 |
| Poisson | 22.00171 | 484.075 |

## Podsumowanie