

Statystyka i Probabilistyka - Projekt 2

Należy przygotować model regresji liniowej prostej dla wybranych przez Państwa danych oraz przeprowadzić diagnostykę modelu. W sprawozdaniu powinny znaleźć się następujące rzeczy:

1. Dokładny opis wykorzystanych danych oraz ich źródło. Dane powinny posiadać przynajmniej 30 obserwacji (zalecane jest dużo więcej).
2. Krótki opis teoretyczny modelu regresji liniowej.
3. Zbudowanie modelu i jego analiza w R. Według pliku, który podesłałem na ostatnich zajęciach. Zbudowanie modelu funkcja `lm`. Interpretacja modelu. Interpretacja wartości z `summary(model)`. Sprawdzenie założeń (normalność, korelacja, autokorelacja reszt, homoskedastyczność itd, wyznaczenie zbioru dla którego model ma sens itd.) Wykresy diagnostyczne z interpretacją, wartości odstające i wpływowe (boxplot, odległość cooke'a). Zbudowanie nowego modelu po usunięciu wartości wpływowych. Na koniec jeszcze Krosvalidacja. Po usunięciu wartości wpływowych dzielimy nasz nowy zbiór na 10 podzbiorów i tworzymy 10 model. Uczymy na 90% zbioru a na 10% testujemy. Liczymy dla obu metrykę R^2 i RSE. Powtarzamy to dla każdego podzbioru (łącznie 10 modeli) i uśredniamy wynik R^2 i RSE dla zbiorów testowych i treningowych. Sprawdzamy czy mamy do czynienia z przetrenowaniem czy nie. Ważne żeby zbiór danych treningowych i testowych był zbalansowany.
4. Na koniec wnioski i podsumowanie (Czy model jest ok itd)