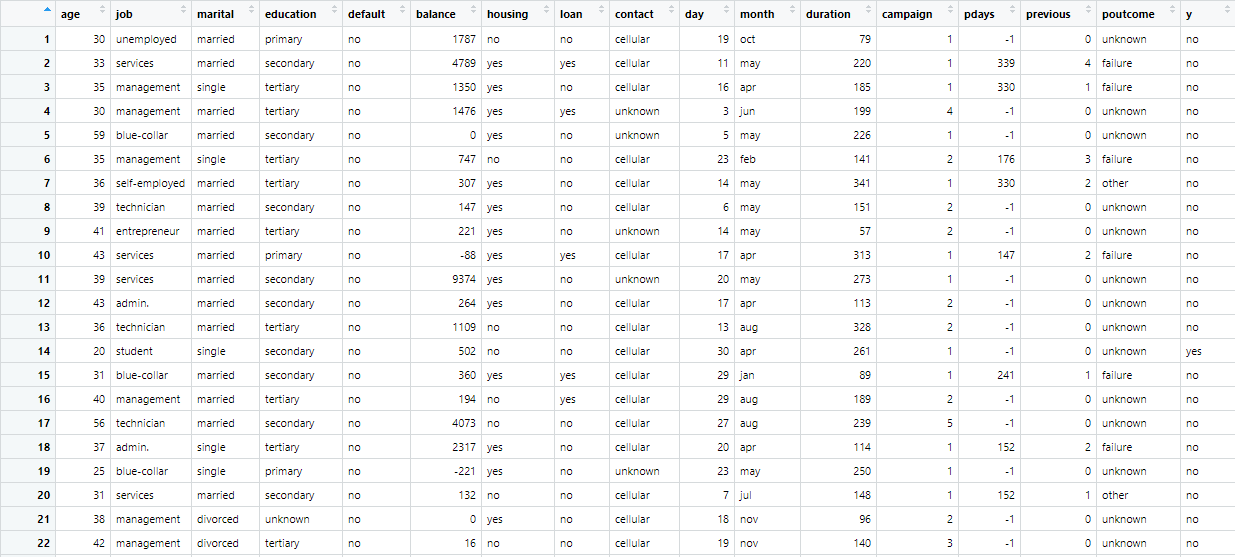
* 1. **Zadanie**

Zadanie polegało na budowie kilku modeli regresji liniowej dla zadanego zestawu danych (wycinek danych przedstawiono poniżej). Ponadto polecono spróbowanie budowy modelu nieliniowego oraz mieszanego i porównanie ich wszystkich ze sobą.



Rys. 1. Fragment danych na podstawie których stworzony został model regresji.

* 1. **Sposób rozwiązania**

Dla zestawu danych przeprowadzono wstępną analizę danych i dokonano korekty „kłopotliwych” zmiennych. Po transformacji danych i doprowadzeniu ich do postaci umożliwiającej przeprowadzenie wymaganych procesów zbudowano model liniowy, nieliniowy oraz mieszany. Na końcu porównano ze sobą modele i wyciągnięto na ich podstawie odpowiednie wnioski.

* 1. **Implementacja w środowisku R**

Zadanie rozpoczęto od wczytania danych poddanych analizie.

****

Po załadowaniu dokonano tzw. samplowania danych, czyli podziału na część estymacyjną oraz walidacyjną. Dzięki takiemu zabiegowi możliwe jest sprawdzenie w jaki sposób model zachowuje się w odniesieniu do jednej z tych dwóch części danych.



W wyniku wykonania powyższego fragmentu kodu do zbioru danych dopisana została dodatkowa kolumna – Sample.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Kolejnym krokiem było przeprowadzenie wstępnej analizy danych. Na podstawie analizy wyciągnięto odpowiednie wnioski, które skutkowały transformacją niektórych danych zawartych w zbiorze pierwotnym. Proces transformacji rozpoczęto od faktoryzacji niektórych danych.



Dokonano również transformacji zmiennej objaśnianej z wartości „prawda”, „fałsz” na odpowiadające im wartości liczbowe „0” oraz „1”. Zabieg ten był konieczny   
do przeprowadzenia regresji logistycznej!



Obraz zawierający stół

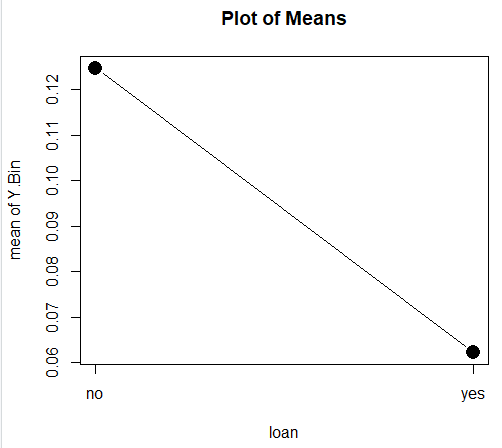
Opis wygenerowany automatycznie

Po przeprowadzeniu transformacji sprawdzono losowo wybrane zależności między różnymi zmiennymi. Celem tego zabiegu miało być sprawdzenie jakiego rodzaju zależności występują między poszczególnymi zmiennymi (liniowe czy nieliniowe).

Obraz zawierający tekst

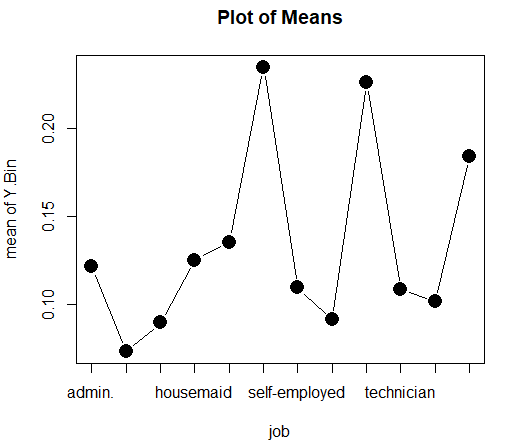
Opis wygenerowany automatycznie

Wykonanie prostych wykresów tego typu pozwoliło na wstępne określenie opisanych wcześniej zależności. Przykładowo badając zależność zmiennej objaśnianej „y”   
od zmiennej „loan”, czyli informacji o posiadaniu osobistego kredytu możemy stwierdzić, że pozostają one w zależności liniowej (Rys. 2).



Rys. 2. Zależność między zmienną objaśnianą, a faktem posiadania kredytu osobistego.

Przykładem zależności liniowej może natomiast stać się następujący wykres (Rys. 3).



Rys. 3. Zależność między zmienną objaśnianą, a rodzajem wykonywanej pracy zawodowej.

Posiadając podstawową wiedzę na temat analizowanego zbioru danych oraz po przeprowadzeniu odpowiednich transformacji mogliśmy przejść do stworzenia pierwszego modelu – uogólnionego modelu liniowego.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Podsumowanie modelu wygląda następująco:

Obraz zawierający stół

Opis wygenerowany automatycznie

Jak nie trudno zauważyć, model liniowy posiada bardzo wysoki współczynnik McFaddena co świadczy o jego dobrej jakości. Niemniej jednak widzimy również, że występuje   
w nim duża liczba zmiennych nieistotnych. W związku z tym dokonano redukcji zmiennych nieistotnych i ponownie przeprowadzono podsumowanie modelu.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Jak widać redukcja zmiennych delikatnie obniżyła jakość modelu, jednakże dalej kierując się współczynnikiem McFaddena, model dalej charakteryzuje się wysoką jakością.

W celu budowy modelu nieliniowego należało usunąć ze zbioru wartości ujemne oraz wykonać logarytmizowanie zmiennych.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Zlogarytmizowanie zmiennych pozwoliło na budowę modelu nieliniowego.



Posumowanie modelu wyszło następująco:

Obraz zawierający stół

Opis wygenerowany automatycznie

Podsumowanie modelu w sposób oczywisty pokazuje wyższość modelu nieliniowego nad liniowym. W przypadku tego modelu współczynnik McFaddena osiąga wartość aż 0.37. Redukcja zmiennych nieistotnych pozwoliła osiągnąć wynik przestawiony poniżej.

Obraz zawierający tekst

Opis wygenerowany automatycznie

W przypadku tego modelu redukcja zmiennych nieco bardziej pogorszyła model pierwotny. Niemniej jednak, dalej pozostaje on na wysokim poziomie!.

Ostatnim krokiem była budowa modelu mieszanego, gdzie w zależności od sprawdzonych wcześniej zależności między zmiennymi wykorzystano zarówno elementy modelu liniowego, jak również nieliniowego.



Wynikiem budowy powyższego modelu było następujące podsumowanie:

Obraz zawierający stół

Opis wygenerowany automatycznie

Jak widać model mieszany przed redukcją osiągnął jakość bardzo zbliżoną do jakości modelu nieliniowego. Po redukcji zmiennych nieistotnych osiągnęliśmy natomiast następujące wyniki:

Obraz zawierający tekst

Opis wygenerowany automatycznie

Co ciekawe, model zredukowany osiągnął wartość identyczną z modelem zredukowanym modelu nieliniowego.

* 1. **Wnioski**

Budowanie modeli regresji w środowisku R jest bardzo przejrzyste i proste   
w obsłudze. Analizując otrzymane w trakcie pracy wyniki możemy stwierdzić kilka rzeczy. Po pierwsze, dla zadanego zestawu danych lepszej jakości okazały się modele nieliniowy i mieszane. Wynika to bezpośrednio z zależności pomiędzy poszczególnymi zmiennymi. Zadanie w sposób klarowny obrazuje jak istotnym elementem podczas tworzenia modeli regresji jest wstępna analiza danych.