Lista 1
Zaawansowane modele liniowe
Regresja logistyczna

Analiza danych

1. Zaimportuj do R dane "*Lista_1.csv*".
   Powyższy zbiór danych opisuje relacje między p-stwami przyjęcia na studia (success) a wynikami z testów rachunkowych (numeracy) i poziomu niepewności (anxiety).

2. Narysuj boxploty dla zmiennej "numeracy" w rozbiciu na grupę przyjętych/nieprzyjętych osób ($boxplot(numeracy \sim success)$). Opisz wnioski.

3. Wykonaj poprzedni punkt dla zmiennej "anxiety".

4. Skonstruuj model regresji logistycznej dla powyższych danych. W szczególności:

   - Podaj estymatory parametrów i wyniki testów istotności,
   - Wyznacz przewidywane p-stwo sukcesu u studenta, którego anxiety=13 a numeracy=10,
   - Wyrysuj krzywą ROC dla dopasowanego modelu statystycznego.

5. Powtórz powyższe ćwiczenie dla różnych funkcji linkujących (probit, cauchit, cloglog) i oceń która z funkcji linkujących daje najlepsze dopasowanie modelu do danych. Porównaj krzywe ROC dla modeli z różnymi funkcjami linkującymi.

6. Skoncentrujemy się obecnie na modelu z funkcją linkującą "logit".

   - Wyznacz estymator macierzy kowariancji wektora estymatorów parametrów w modelu regresji logistycznej. Następnie, porównaj wartości na przekątnej z estymatorami odchyleń standardowych zwracanych przez R.
   - Przetestuj jedną hipotezę, że obie zmienne objaśniające nie mają wpływu na zmienną odpowiedzi.

- Przetestuj hipotezę, że rozkład danych jest zgodny z założonym modelem.
- Podaj definicję parametru "epsilon" i jego wartość domyślną. Wykonaj ponownie obliczenia stosując wartości epsilon ze zbioru: $10^{-1}, 10^{-2}, 10^{-3}$ i 10^{-6}. Porównaj liczbę iteracji i wartości estymatorów poszczególnych parametrów.

Symulacje

1. Wygeneruj macierz X wymiaru $n = 400, p = 3$, której elementy są zmiennymi losowymi z rozkładu $N(0, \sigma^2 = 1/400)$ (pamiętaj, że funkcja $rnorm()$ wymaga podania parametru σ(nie σ^2!)). Załóżmy, że binarny wektor odpowiedzi jest wygenerowany zgodnie z modelem regresji logistycznej z wektorem $\beta = (3, 3, 3)$. Wyznacz macierz informacji Fishera w punkcie β i asymptotyczną macierz kowariancji estymatorów największej wiarogodności. Nastepnie wygeneruj 1000 replikacji wektor odpowiedzi zgodnie z powyższym modelem i na podstawie każdej replikacji wyznacz estymator wektora β. W tym celu skorzystaj z funkcji $glm()$, przy czym wyłącz z modelu Intercept, gdyż nie korzystamy z niego przy generowaniu danych ($glm(y \sim X - 1, ...)$). Na podstawie uzyskanego zbioru estymatorów:

   - Narysuj histogramy estymatorów $\hat{\beta}_1, \hat{\beta}_2$ i $\hat{\beta}_3$ i porównaj z ich rozkładami asymptotycznymi.
   - Wyestymuj obciążenie estymatorów $\hat{\beta}_1, \hat{\beta}_2$ i $\hat{\beta}_3$
   - Wyestymuj macierz kowariancji wektora estymatorów ($\hat{\beta}_1, \hat{\beta}_2$ i $\hat{\beta}_3$) i porównaj z asymptotyczną macierzą kowariancji.

2. **Wpływ liczby obserwacji n**.
   Doswiadczenie powtórz w przypadku gdy n=100.

3. **Wpływ korelacji miedzy regresorami**.
   Punkt 1 powtórz w przypadku gdy wiersze macierzy X sa niezależnymi wektorami losowymi z wielowymiarowego rozkładu normalnego $N(0, \Sigma)$ z macierza kowariancji $\Sigma = \frac{1}{n}S$, gdzie $S_{ii} = 1$, a dla $i \neq j$, $S_{ij} = 0.3$.

4. **Wpływ liczby regresorów**.
   Punkt 1 powtórz w przypadku, gdy elementy X sa niezależne a p=20.

5. Porównaj wyniki i opisz wnioski.