Logistic Regression

Tasks (Lab 3):

1. Interpretation of the model, testing.

   Dataset *SAheart.data* (South African Heart Disease) contains information about patients with heart diseases. Variable **chd** indicates if heart attack has occurred (value 1) or not (value 0). Detailed description of the dataset you can find in file *SAheart.info*.

   - Fit logistic regression model, treating **chd** as response variable.
   - Which variables are significant in the full model?
   - How to interpret the coeficient correspondg to variable **Age**?
   - What is the predicted probability of the heart attack for the first patient in the dataset?

2. Problem of linearly separable classes.

   Dataset *earthquake.txt.* corresponds to problem of prediction of seismic shocks (volcanic eruptions and nuclear explosions) (variable **popn**) based on two variables: **body** (deep wave magnitude) and **surface** (surface wave magnitude).

   - Make scatterplot for variables **body** and **surface**. Mark classes corresponding to observations.
   - Fit logistic model and perform Wald test (test for significance of the coefficients). How to explain the fact that the test indicates that both variables are not significant in the full model?

3. Simulation example (**scored task**).

   - Generate data from logistic model:

     $$y_i \sim Bern(p_i),$$

     where

     $$p_i = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})]},$$

     for $i = 1, \ldots, n$, $x_{1,i}$, $x_{2,i} \sim N(0, 1)$, $n = 50$. Parameters: $\beta_0 = 0.5$, $\beta_1 = \beta_2 = 1$. Fit logistic model and calculate the estimators of the coefficients. Repeat the experiment $L = 100$ times and compute the MSE (mean squared error):

     $$MSE = E(||\hat{\beta} - \beta||^2),$$

     where $||\cdot||$ is Euclidean norm and $\beta = (\beta_0, \beta_1, \beta_2, beta_3)$ is vector of true parameters.
   - Repeat the experiment for $n = 50, 60, 70, 80, 90, 100, 200, 300 \ldots, 1000$ and make a plot showing how MSE depends on n. Save the results in the file **LogisticSimulatedData1.pdf**.