Course: Data Analysis (task № 7)

Student’s Name and Surname **Iurii Aksenov**

Please do the tasks below.

Open **Warranty.dta** file.

The aim of analysis is to find out what circumstances encourage customers to purchase extended warranties after a major appliance purchase. The response variable is an indicator of whether or not a warranty is purchased (Bought). The predictor variables are:

* Customer’s gender (Gender)
* Customer’s age (Age)
* Whether a gift is offered with the warranty or not (Gift)
* Price of the appliance (Price100)
* Customer’s race (Race)

Use the binary logistic regression to analyze the data.

1. Specify the regression equation (the linear part of the formula).

**logit Bought i.Gender i.Gift Age i.Race Price100**



If all parameter is 0, then the bought of warranty is -8.244.

1. Assess the goodness-of-fit of the model and interpret the results of the analysis.

**estat gof**

The model is significantly better than the baseline model because Prob = 0.0 <0.05





The goodness-of-fit test is positive.

H0 there is no difference between real and predicted values

H1 there is difference

P = 1 > 0.05 => reject H1.

There is no significance difference between real and predicted values of dependent variable

1. Interpret the influence of any predictor variable on the dependent variable using Exp(b).

**logistic Bought i.Gender i.Gift Age i.Race Price100**



For males the probability to not buy the warranty is 0.98 odds.

For the sales with the gift offering the probability to buy warranty is 15 times greater than without offering the gift.

African American in 0.93 odd don’t buy the warranty in comparison with white.

Hispanic more in 13 odd times buy warranty in comparison with white.

One unit increase in age will increase the probability to buy warranty in 1.0957 times.

One unit increase in price will increase the probability to buy warranty in 1.0956 times.

4) Which gradients are statistically significant?



All P values are more than 0.05, so all observing gradients are statistically significant.

5) What is the percentage of correctly predicted cases by the model?

**estat classification**



The percentage of correctly predicted cases by the model is 94.0 %

6) Do the diagnostics of the model.

- Are the residuals normally distributed?

// create residuals

predict r, rstandard

// check whether the residuals normal standardized

swilk r

// visualizing the residuals

kdensity r, normal



H1 says that the distribution of residual is significantly different from normal distribution.

The p-value<0.05 => we accept H1=> residual are not normally distribute.



By the graph we can see that the residual is not normally distributed.

- Are there any outliers? If yes, how many?

**calculate the frequencies of residuals where r >3 | r<-3**

**tabulate r if r >3 | r<-3**

**summarize r, detail**





We have one outlier with the value 4.19.

- Test the multicollinearity.

To create dummy variable from the categorical variable Race

**tabulate Race, generate(Race\_dummy)**

**corr Bought Gender Gift Age Race\_dummy1 Race\_dummy2 Race\_dummy3 Race\_dummy4 Price100**



There are no multicollinearity variables because all values of the correlation matrix are less than 0.75.