Course: Data Analysis

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

**Regression analysis**

Do the multiple regression analysis using the following variables from **data\_games.dta**:

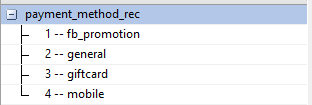
Dependent variable: payment

Independent variables: payment\_type, payment\_method, crystalls\_balance\_before\_buy, crystalls\_bought

***Describe the model:***

1. **Specify the regression equation.**

Firstly, we have categorical variable **payment\_method** with more than 2 values.



So, regression cannot be used with such variable => we will create dummy variables.



If all parameters are 0, then the payment is -609.

The average **payment** of **general** payment method bigger in 256 than the average **payment** in **fb\_promotion**.

1. **Assess the goodness-of-fit of the model (R-square, significance of the model);**



The regression model is statistically significant because the p-value < 0.05. The model explains 73.9% of payment variation.

1. **Which independent variables significantly influence the dependent variable (explain your answer)?**



We have that all of our variables are significantly correlate with predictor because their p-values < 0.05.

1. **Describe the relationship between each independent and the dependent variable.**

There relationships between each independent and dependent variables because all have p-values < 0.05.

There is negative significant relationship between crystalls\_balance\_before\_buy. One unit increase in crystalls\_balance\_before\_buy will decrees the payment by - 0.606.

There is positive significant relationship between crystals\_boughts and payment. One unit increase crystals\_boughts will increase payment by 16.17.

The payment of regular paymeny\_type is significantly higher than the offer paymeny\_type. The average difference is 395.

The general payment\_method is significantly higher than the fb\_promotion payment\_method. The average difference is 256.

The giftcard payment\_method is significantly lower than the fb\_promotion payment\_method. The average difference is -1920.

The mobile payment\_method is significantly lower than the fb\_promotion payment\_method. The average difference is -342.

Let’s calculate the beta to compare the strength of relationships.



As for beta the most predictor is payment\_type than crystals\_bought.

***Do the diagnostics of the model:***

1. **Are the residuals normally distributed?**



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

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



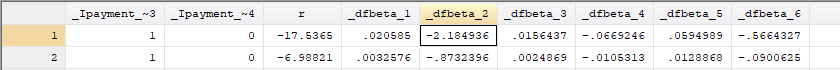
1. **Are there any outliers (standardized residuals greater than 3, or less than -3)?**

We have outliers. The amount of outliers is 1778.

As for Influential cases.



We have Influential case because the second dfbeta\_2 more the |1|.



1. Test the multicollinearity and heteroscedasticity.

**Multicollearity**



There is no problem because all VIF less than 10 => three is no Multicollearity

**Heteroscedasticity**



H0 no relationship between residuals

There is a heteroscedasticity because the p-value < 0.05– relationship between residuals and the fitted values of the payment.

****

Please send this MS Word file with answers and the do-file to amelikyan@hse.ru from your personal e-mail.