



Insurance Claims Case Study: Linear vs Gamma Generalized Linear Model Regression

NovaMechanics Ltd

Introduction

The goal of this study is to perform a statistical analysis of health insurance charges data and to highlight meaningful relationships within the data, using the Isalos Analytics Platform. The dataset used in this study comprises both continuous and categorical variables describing the insured party's characteristics and way of life (Table 1), and is available in the following address: [Health Insurance Claims](#).

Table 1. Description of variables in the insurance claims dataset.

Variable	Description
Age	Continuous
Sex	Male, Female
BMI	Continuous
Children	Continuous
Smoker	Yes, No
Region	Southeast, Southwest, Northeast, Northwest
Charges	Target variable. Continuous

The analysis was conducted by testing two Generalized Linear Models (GLMs)-the linear model and the gamma model with a log link- using different distributions and link functions to identify the best-fitting model based on commonly used Goodness-of-Fit metrics. Specifically, for the right-skewed insurance charges target, the linear model is expected to perform poorly, while the gamma GLM is typically more suitable. For further details on the use of different GLMs, please refer to the Isalos GLM documentation: [Introduction | Isalos Analytics Platform Docs](#).

GLM Regression Analysis in Isalos

Step 1: Import data from file

The dataset can be imported on Isalos by right clicking on the input spreadsheet and selecting **Import from file** (Fig. 1, 2).

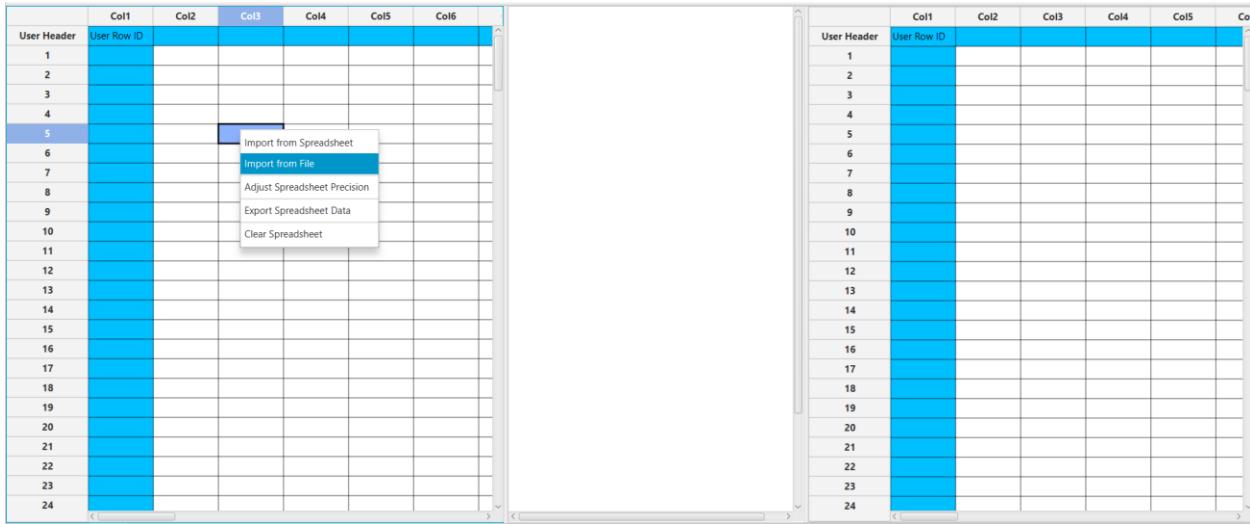


Figure 1. Import data from file by right-clicking on the input spreadsheet and selecting the **Import from file** option.

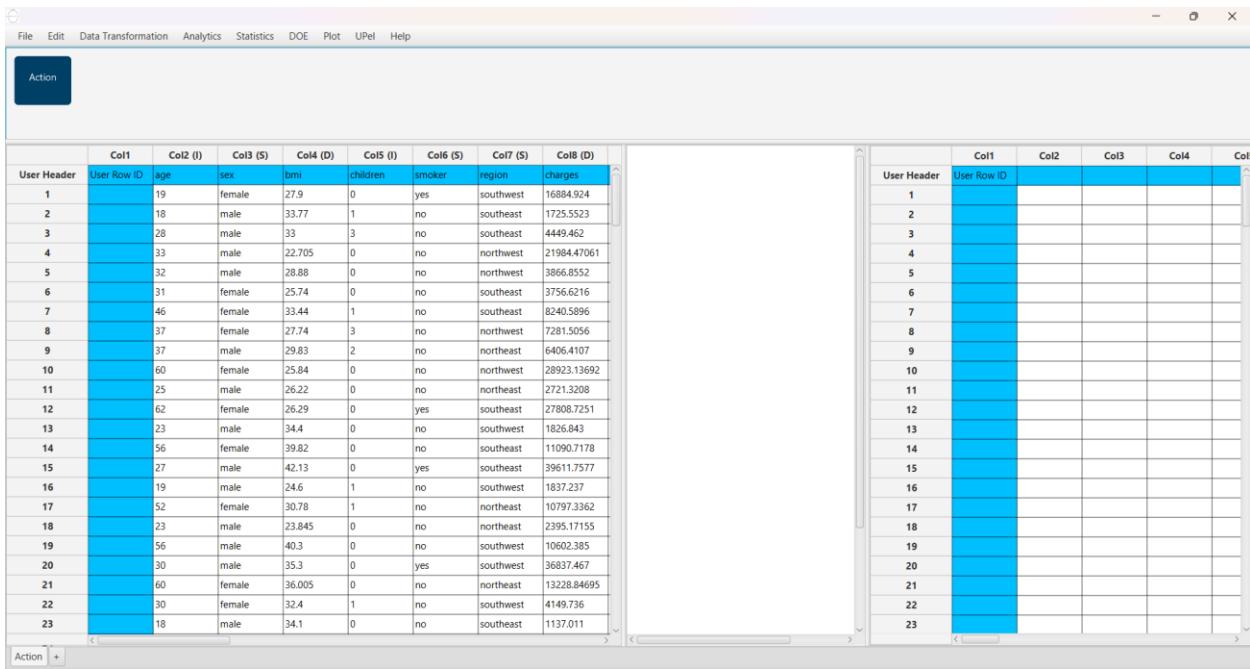


Figure 2. Inspect the imported data on the right-hand spreadsheet.

Step 2: Perform the data analysis

By navigating the tools ribbon and selecting **Analytics → Regression → Statistical fitting → Generalized Linear Models** (Fig. 3), the Generalized Linear Models Regression window pops-up (Fig. 4).

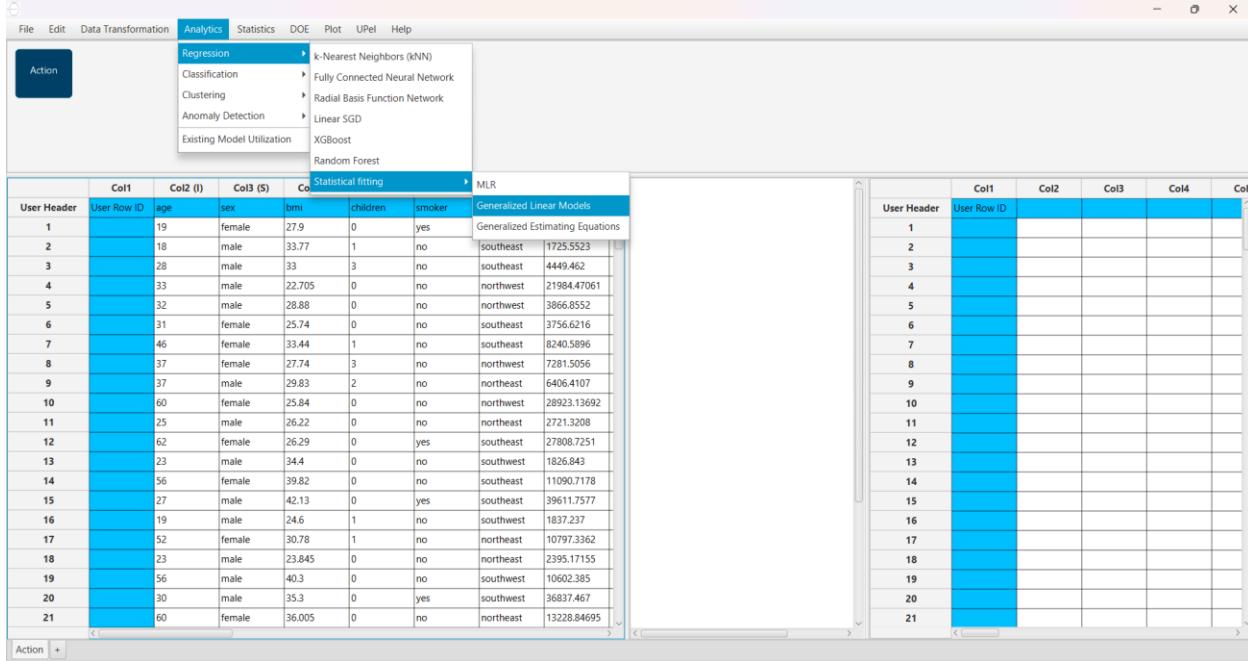


Figure 3. Navigate the tool bar to select the Regression GLMs option.

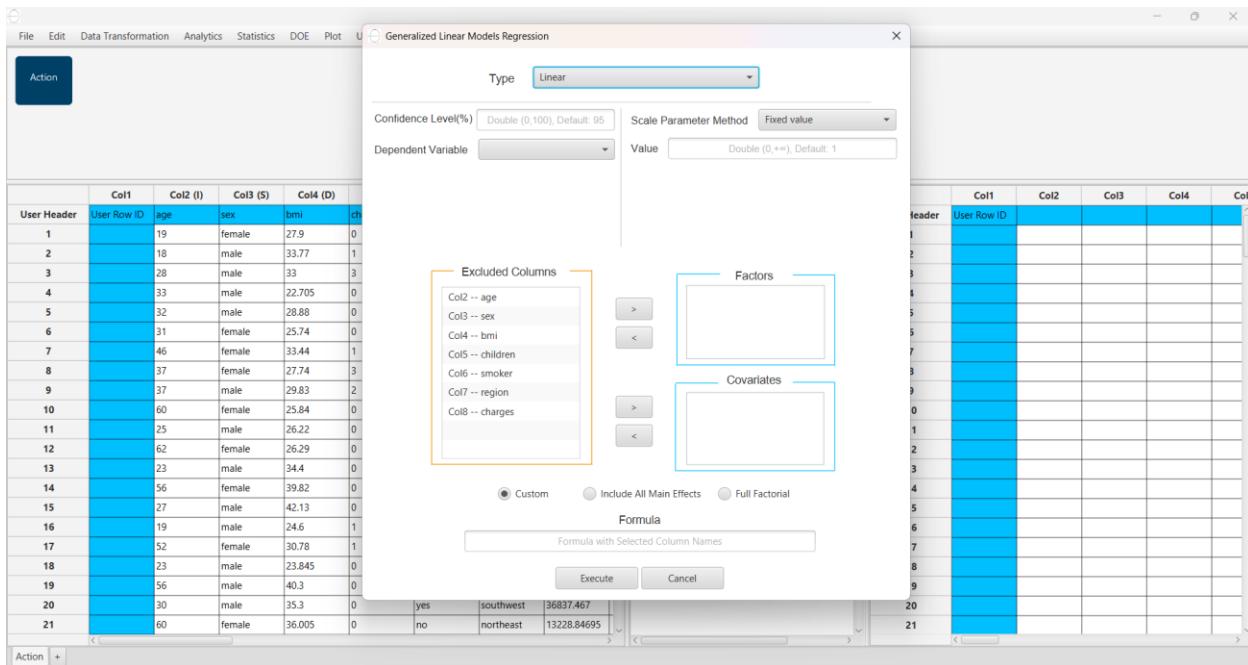


Figure 4. The Generalized Linear Models Regression window pops-up.

In the Generalized Linear Models Regression window, the GLM type-in this case the linear model- can be selected from the drop-down menu (Fig. 5A). The model's target variable (charges) and configuration parameters can be set as shown in Fig. 5B, while the included factors (categorical variables: Sex, Smoker, Region) and covariates (continuous variables: Age, BMI, Children) are selected as shown in Fig. 5C. For this analysis the model parameters are set to the default Isalos values, and the option **Include All Main Effects** is selected (Fig. 5D). The model is fitted to the data by clicking **Execute** (Fig. 5E).

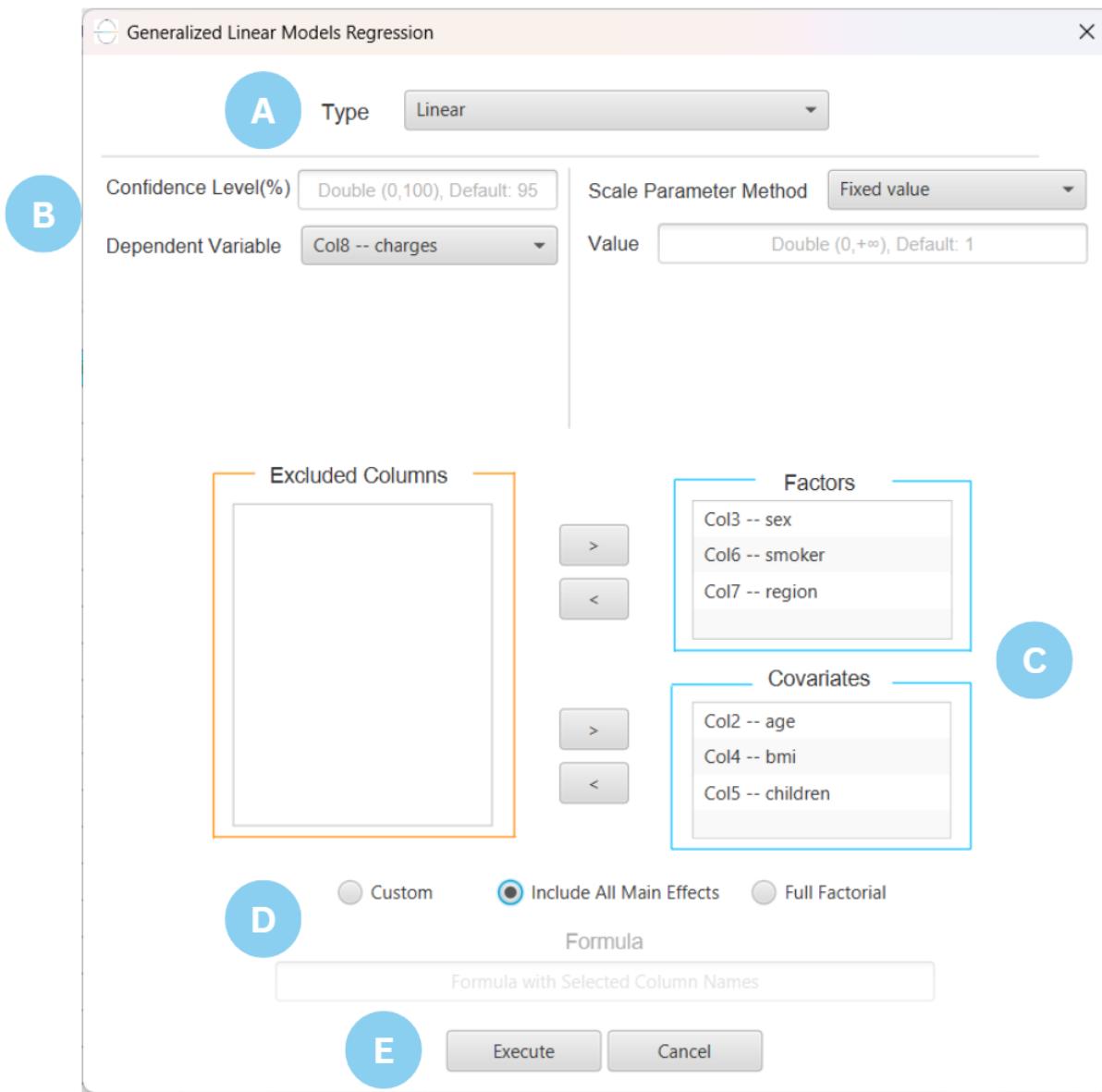


Figure 5. Linear GLM configuration. A. GLM type, B. Model parameters and target variable, C. Selected factors and covariates, D. Analysis type, E. **Execute** button.

The same steps can be followed to apply the gamma GLM with a log link, with the addition of specifying more parameter values (Fig. 6B). In this case, the model parameters were also set to the default Isalos values.



Figure 6. Gamma GLM configuration. A. GLM type, B. Model parameters and target variable, C. Selected factors and covariates, D. Analysis type, E. **Execute** button.

Step 3: Results

Upon clicking **Execute** the modelling results appear on the output spreadsheet (Fig. 7, 8). By right clicking on the output spreadsheet and choosing **Export Spreadsheet Data** the results can be downloaded in XLSX or CSV format.

	Col1	Col2 (D)	Col3 (D)	Col4	Col5 (S)	Col6 (S)	Col7 (S)	Col8 (S)	Col9 (S)	Col10 (S)	Col11 (S)	Col12 (S)
User Header	User Row ID	charges	Prediction									
1		16884.924	25293.7130284		Goodness of Fit							
2		1725.5523	3448.6028343			Value						
3		4449.462	6706.9884907		Deviance	48839532843.921844						
4		21984.47061	3754.8301630		Scaled Deviance	48839532843.921844						
5		3866.8552	5592.4933865		Pearson Chi-Square	48839532843.921844						
6		3756.6216	3719.8257990		Scaled Pearson Chi-Square	48839532843.921844						
7		8240.5896	10659.9612251		Log Likelihood	-24419767651.50068						
8		7281.5056	8047.9106069		Akaike's Information Criterion (AIC)	48839535321.00136						
9		6406.4107	8502.9739198		Finite Sample Corrected AIC (AICC)	48839535321.1369						
10		28923.13692	11884.6375180		Bayesian Information Criterion (BIC)	48839535367.79174						
11		2721.3208	3245.2082315		Consistent AIC (CAIC)	48839535376.79174						
12		27808.7251	35717.4636691									
13		1826.843	4546.0469857									
14		11090.7178	14917.0784393		Parameter Estimates							
15		39611.7577	31969.0012761		Variable	Coefficient	Std. Error	Lower CI	Upper CI	Test Statistic	df	p-value
16		1837.237	670.0262753		intercept	10818.6306150	0.1768058	10818.2840821	10818.9771480	3744136149.386466	1	0.0
17		10797.3362	12333.8668031		age	256.8563525	0.0019628	256.8525055	256.8601996	17124473620.848646	1	0.0
18		2395.17155	1925.9110741		sex_female	131.3143594	0.0549224	131.2067134	131.4220054	5716429.0666148	1	0.0

Figure 7. Linear GLM results.

	Col1	Col2 (D)	Col3 (D)	Col4	Col5 (S)	Col6 (S)	Col7 (S)	Col8 (S)	Col9 (S)	Col10 (S)	Col11 (S)	Col12 (S)
User Header	User Row ID	charges	Prediction									
1		16884.924	15984.2071387		Goodness of Fit							
2		1725.5523	3882.4963767			Value						
3		4449.462	6051.6256365		Deviance	337.7266376						
4		21984.47061	5100.6069443		Scaled Deviance	337.7266376						
5		3866.8552	5408.3024369		Pearson Chi-Square	620.5798055						
6		3756.6216	4895.2435848		Scaled Pearson Chi-Square	620.5798055						
7		8240.5896	9122.6248623		Log Likelihood	-13680.8686988						
8		7281.5056	8369.7031430		Akaike's Information Criterion (AIC)	27379.7373976						
9		6406.4107	7930.8582316		Finite Sample Corrected AIC (AICC)	27379.8729397						
10		28923.13692	12230.5797347		Bayesian Information Criterion (BIC)	27426.5277787						
11		2721.3208	4516.6705371		Consistent AIC (CAIC)	27435.5277787						
12		27808.7251	53738.9627463									
13		1826.843	4139.3915839									
14		11090.7178	12219.9974233		Parameter Estimates							
15		39611.7577	23299.9592983		Variable	Coefficient	Std. Error	Lower CI	Upper CI	Test Statistic	df	p-value
16		1837.237	3496.4820270		intercept	8.6841075	0.1835191	8.3244166	9.0437984	2239.1731392	1	0.0
17		10797.3362	12020.9790179		age	0.0286379	0.0019802	0.0247569	0.0325190	209.1621363	1	0.0
18		2395.17155	4124.5442370		sex_female	0.0570778	0.0550479	-0.0508141	0.1649697	1.0751103	1	0.2997940
19		10602.385	11575.9585691		bmi	0.0141237	0.0048387	0.0046400	0.0236073	8.5199756	1	0.0035127

Figure 8. Gamma GLM results.

Discussion and Model Selection

The analysis results include the Predictions for the target values, the Goodness-of-Fit statistics, and the Parameter Estimates table. Model evaluation and selection should be based on the goodness-of-fit metrics. In brief, the deviance metric indicates how well a model fits the data, with lower values reflecting higher predictive accuracy. The chi-square metrics measure the discrepancy between observed and expected values, with smaller values also indicating a better model fit. Finally, the Akaike and Bayesian Information Criteria are informative only when compared across models, with the model showing the lowest values considered to perform better.

As expected for this type of data, the gamma GLM Goodness-of-Fit metrics are significantly better than those of the linear model, with lower deviance, chi-square, and Akaike and Bayesian Information Criteria values (Table 2). Therefore, the results indicate that the linear model is unsuitable for analysing this dataset, where the target variable exhibits a right-skewed distribution.

Table 2. Comparison of Goodness-of-Fit Statistics for linear and gamma GLMs.

Metric	Linear GLM	Gamma GLM
Deviance	4.88E+10	3.38E+02
Scaled Deviance	4.88E+10	3.38E+02
Pearson Chi-Square	4.88E+10	6.21E+02
Scaled Pearson Chi-Square	4.88E+10	6.21E+02
Log Likelihood	-2.44E+10	-1.37E+04
Akaike's Information Criterion (AIC)	4.88E+10	2.74E+04
Finite Sample Corrected AIC (AICC)	4.88E+10	2.74E+04
Bayesian Information Criterion (BIC)	4.88E+10	2.74E+04
Consistent AIC (CAIC)	4.88E+10	2.74E+04

The final step in the data analysis case study is interpreting the results of the selected model. The Parameter Estimates table presents the coefficients of the predictive formula, along with their standard errors, confidence intervals (Lower CI and Upper CI), test statistics, degrees of freedom (df), and p-values (Table 3). Considering that in this analysis the confidence level was set to 95%, only variables with a p-value lower than 0.05 are considered statistically significant. Therefore, in the predictive formula (Eq. 1) only the Intercept, Age, BMI, Children and Smoker_no variables should be included.

Table 3. Parameter Estimates table for the gamma GLM.

Variable	Coefficient	Std. Error	Lower CI	Upper CI	Test Statistic	df	p-value
intercept	8.684	0.184	8.324	9.044	2239.173	1	0.000
age	0.029	0.002	0.025	0.033	209.162	1	0.000
sex_female	0.057	0.055	-0.051	0.165	1.075	1	0.300
bmi	0.014	0.005	0.005	0.024	8.520	1	0.004
children	0.084	0.023	0.039	0.129	13.254	1	0.000
smoker_no	-1.500	0.070	-1.637	-1.363	460.358	1	0.000
region_northeast	0.145	0.079	-0.009	0.300	3.412	1	0.065
region_northwest	0.088	0.079	-0.067	0.242	1.233	1	0.267
region_southeast	0.004	0.077	-0.148	0.155	0.002	1	0.960

Predictive formula:

$$\logCharges = 8.684 + 0.029 * Age + 0.004 * BMI + 0.084 * Children - 1.500 * Smoker_no \text{ (Eq. 1)}$$

Equation 1 shows that the logarithm of health insurance charges is expected to decrease substantially for non-smokers, while it increases with higher BMI, a greater number of children, and, particularly, age.

Model utilization

The selected model can be fitted to unknown data by browsing the tools ribbon and selecting **Analytics → Existing Model Utilization** to predict the health insurance charges for unknown insured parties (Fig. 9). The new dataset must contain all input variable columns to acquire predictions.

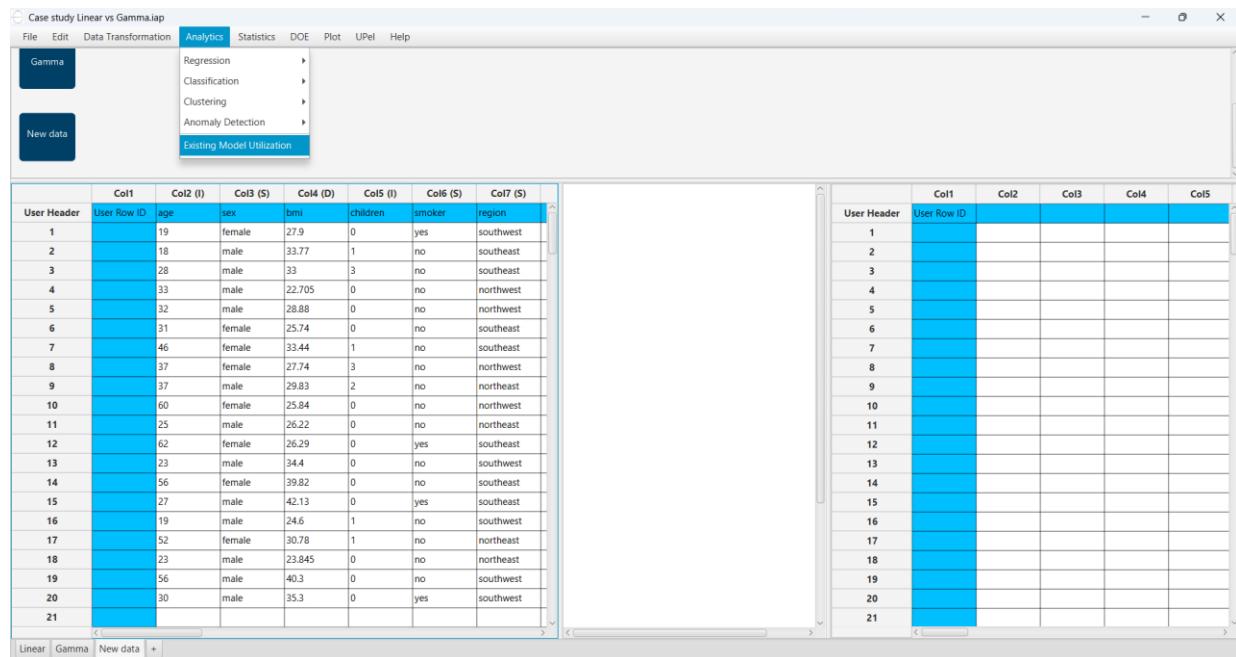


Figure 9. Fit the selected model to new data by selecting **Analytics → Existing Model Utilization**.

In the Existing Model Execution pop-up window, the fitted model-in this case, the gamma GLM- can be selected (Fig. 10A) and utilized to predict the new data by clicking **Execute** (Fig. 10B).

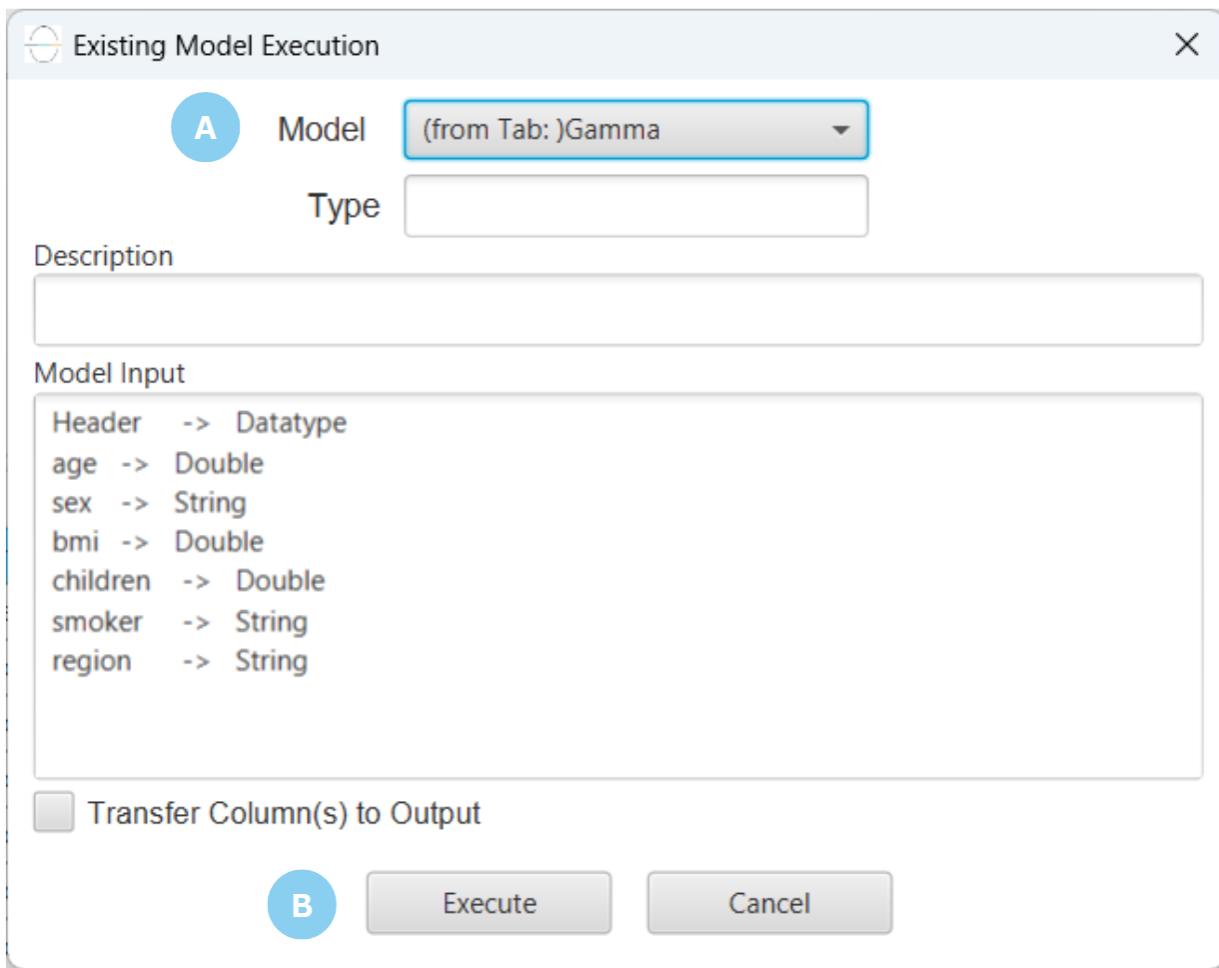


Figure 10. Existing Model Execution pop-up window: A. Existing model selection, B. **Execute** button.

Upon execution the prediction results appear on the output spreadsheet (Fig. 11).

	Col1	Col2 (D)	Col3	Col4	Col5	Col6	Col7
User Header	User Row ID	Prediction					
1		15984.2071387					
2		3882.4963767					
3		6051.6256365					
4		5100.6069443					
5		5408.3024369					
6		4895.2435848					
7		9122.6248623					
8		8369.7031430					
9		7930.8582316					
10		12230.5797347					
11		4516.6705371					
12		53738.9627463					
13		4139.3915839					
14		12219.9974233					
15		23299.9592983					
16		3496.4820270					
17		12020.9790179					
18		4124.5442370					
19		11575.9585691					
20		22966.9416004					
21							
22							

Figure 11. Gamma GLM predictions for unknown data.

Further reading

Rayner, J.C.W., O. Thas, and D.J. Best. 2009. Smooth Tests of Goodness of Fit: Using R. Wiley Series in Probability and Statistics Smooth Tests of Goodness of Fit Using R. Wiley. <https://books.google.gr/books?id=bDUEafBSZ4C>.

Ailobhio, D. T., and J. A. Ikughur. 2024. “A Review of Some Goodness-of-Fit Tests for Logistic Regression Model.” *Asian Journal of Probability and Statistics* 26 (7): 75–85. <https://doi.org/10.9734/ajpas/2024/v26i7631>.