Journal of Modern Applied Statistical Methods

Volume 18 | Issue 1 Article 1

2-13-2020

Regression Modeling and Prediction by Individual Observations versus Frequency

Stan Lipovetsky
GfK North America, stan.lipovetsky@gmail.com

Recommended Citation

Lipovetsky, S. (2019). Regression modeling and prediction by individual observations versus frequency. Journal of Modern Applied Statistical Methods, 18(1), eP2692. doi: 10.22237/jmasm/1556669100

Regression Modeling and Prediction by Individual Observations versus Frequency

Erratum

An earlier version of this article mistakenly replaced ellipses with a capital letter K. This has been corrected.

Journal of Modern Applied Statistical Methods
May 2019, Vol. 18, No. 1, eP2692.
doi: 10.22237/jmasm/1556669100

בס"ד

ISSN 1538 − 9472

Regression Modeling and Prediction by Individual Observations versus Frequency

Stan Lipovetsky
GfK North America
Minneapolis, MN

A regression model built by a dataset could sometimes demonstrate a low quality of fit and poor predictions of individual observations. However, using the frequencies of possible combinations of the predictors and the outcome, the same models with the same parameters may yield a high quality of fit and precise predictions for the frequencies of the outcome occurrence. Linear and logistical regressions are used to make an explicit exposition of the results of regression modeling and prediction.

Keywords: Multiple regression, modeling, prediction, linear and logistic regression, meaning and interpretation of results, *p*- and *D*-value for significance estimation

Introduction

Regression modeling is, probably, the main tool of applied statistics used for various aims of estimation and prediction. Many works are devoted to numerous models and their specific characteristics useful in practical regression modeling (Kendall & Stuart, 1973; McCullagh & Nelder, 1999; Train, 2003; Izenman, 2008; Andersen & Skovgaard, 2010; Grafarend & Awange, 2012; Härdle & Simar, 2012; Hilbe & Robinson, 2013; Kuhn & Johnson, 2013; Wilson & Lorenz, 2015; Lipovetsky & Conklin, 2001, 2010, 2015). In spite of apparently well-studied area of regressions' features and applications, researchers and practitioners can encounter different phenomena not noticed previously. For instance, a model built by a dataset could show a low quality of fit and poor predictions of individual observations; however, usage of the frequencies of possible combinations of the predictors and the outcome yields the same model with the same parameters but with a high quality of fit and precise predictions.

doi: 10.22237/jmasm/1556669100 | Accepted: September 18, 2017; Published: February 13, 2020.
Correspondence: Stan Lipovetsky, stan.lipovetsky@gmail.com

The aim of this study is to analyze reasons of such seemingly paradoxical results. Linear and logistical models are used for illustration of the regressions results (Lipovetsky, 2012, 2013, 2018). Meaningful application of regressions are necessary for practical needs and helps managers and decision makers to improve understanding and real use of statistical models. Besides this main topic, the work also considers tests with the *p*-value and *D*-value in relations to the predictions by regression models.

Modeling and Prediction by Observations and by Their Frequencies

Consider some main relations of regression modeling needed for further description of the problem. Consider a dependent variable y and predictors x_j (j = 1, 2,…, n; number of variables), and there are observations by them, y_i and x_{ij} (i = 1, 2,…, N; base size). A multiple linear regression can be presented in the model

$$y = a_0 + a_1 x_1 + \ldots + a_n x_n + e, \tag{1}$$

where a_j are the model parameters and e denotes deviation from the model, or the error term. Minimizing the objective of squared errors

$$S^2 = \sum_{i=1}^{N} \left(y_i - a_0 - a_1 x_{i1} - \ldots - a_n x_{in} \right)^2 \tag{2}$$

yields solutions for parameters of the ordinary least squares (OLS) regression. The minimum of the OLS criterion (2) is called residual sum of squares S^2_{res}. The quality of the data fit is convenient to estimate via the coefficient of multiple determination R^2 defined as

$$R^2 = 1 - \frac{S^2_{\text{res}}}{S^2_{\text{orig}}}, \tag{3}$$

where S^2_{res} and S^2_{orig} are the residual and original sum of squares of the dependent variable relatively the regression predictions and the mean level, respectively. The coefficient R^2 has values from zero to one, for the worst and the best quality of fit, respectively.

Sometimes there could be the same set of predictors' values in different groups of observations. For instance, consider a simple case with three binary predictors, $n = 3$, so there is a maximum $M = 8$ of possible combinations, or cells of their unique combined values (cells can be numerated as $m = 1, 2,\ldots, M$; total number of cells). For each of these cells we find the number of incidents $N_m(y = 1)$ and total observations N_m, so their quotient yields the mean value of y, or frequency f of the event $y = 1$ in each m^{th} cell:

$$f_m = \frac{N_m\left(y=1\right)}{N_m}. \tag{4}$$

The frequencies f can be used as the outcome values in place of y in the linear model (1) by all N observations:

$$f = b_0 + b_1 x_1 + \ldots + b_n x_n + \delta, \tag{5}$$

where b_j denote parameters estimated by this model for frequency and δ are deviations. OLS minimization (2) can be used for finding the model (5), but instead of using the same f values within a cell, we can collapse N rows of the data matrix into M rows of different cells and use the weights N_m of number of observations in each cell in the weighted least squares (WLS):

$$S^2 = \sum_{m=1}^{M} N_m \left(f_m - b_0 - b_1 x_{m1} - \ldots - b_n x_{mn} \right)^2 . \tag{6}$$

The coefficient of multiple determination for this linear model can be calculated as in (3).

For a binary dependent variable y with the outcome 0 and 1 the logistic regression is commonly applied, with the probability of the event defined by the logit model:

$$p = \frac{1}{1+\exp\left(-\left(c_0 + c_1 x_1 + \ldots + c_n x_n\right)\right)}. \tag{7}$$

Parameters c_j of this model are found in the maximum likelihood (ML) objective for the binomial distribution

$$ML = \prod_{i=1}^{N} p_i^{y_i} \left(1 - p_i\right)^{1-y_i}, \tag{8}$$

The quality of fit in the logistic regression (7) can be evaluated by the so-called pseudo-R^2 defined via the residual and null deviances (those proportional to the logarithm of ML objectives for the models with and without predictors, respectively):

$$R^2 = 1 - \frac{\text{residual deviance}}{\text{null deviance}}, \tag{9}$$

which is constructed in the analogue to the OLS coefficient of multiple determination (3). It is interesting to note that the quotient in (9) corresponds to the percentage of entropy and the maximum possible entropy that defines the measure of efficiency in the information theory (Lipovetsky, 2015).

With the calculated probabilities p (7), the log-odds transformation presents the model (7) in the so-called linear link function

$$\ln \frac{p}{1-p} = c_0 + c_1 x_1 + \ldots + c_n x_n . \tag{10}$$

The relation (10) can also be used for finding the model parameters in the approach proposed and applied for complicated problems of marketing research in Lipovetsky and Conklin (2014) and Lipovetsky (2015). Consider the simple case with three binary predictors when we have $M = 8$ possible cells of their unique combined values. As it is described in (5) we find the frequency f of the event $y = 1$ in each cell and define with them the empirical log-odds values denoted as z

$$z = \ln \frac{f}{1-f} . \tag{11}$$

The values z used at the left-hand side in the linear link (10) for all N observations, so this model can be constructed as a linear regression:

$$z = d_0 + d_1 x_1 + \ldots + d_n x_n . \tag{12}$$

To distinguish parameters c_j estimated by the logit model (7) and by the linear link with log-odds of empirical frequencies (12), the estimated parameters in (12) are denoted as d_j. The model (12) can be constructed in OLS approach, and the coefficients d_j define the logit model (7) so they can be used for prediction by the logistic regression. Instead of using the same z values within each cell we can collapse N rows of the data matrix into M rows of different cells and use weights for cells equal the number of observations in each cell, as it is done in the WLS objective (6).

Numerical Examples

For a clear exposition, consider several numerical examples using datasets taken from real marketing research projects:

Example A

In this dataset, there are three binary variables x_1, x_2, x_3 of advertising, shown or not, and a binary outcome variable y of bought or not (1 or 0, respectively). In a sample of $N = 400$ respondents each one could see maximum one advertising before answering on the purchase interest y about a specific product.

Presented in Table 1 are the coefficients a_j of the linear regression OLS estimation (1)-(2), coefficients b_j of the WLS linear regression (5)-(6) for the frequency model, the logit c_j estimates (7)-(8), and the linear link WLS estimates d_j (11)-(12). The model's quality measures R^2 are shown in the bottom row of Table 1: for the linear estimations (1), (5), and linear link regression (12) the coefficient R^2 (3) is used, and for the logit regression (7) the pseudo-R^2 (9) is used.

Table 1. Example A: Parameters of the linear, logit, and linear link regressions

	Linear regression model		Logistic regression model	
	OLS by observations ***a*** **(1)**	WLS by cells ***b*** **(5)**	ML by observations ***c*** **(7)**	Linear link by cells ***d*** **(12)**
Intercept	0.01923	0.01923	−3.93183	−3.93183
x_1	0.02979	0.02979	0.96655	0.96655
x_2	0.00376	0.00376	0.18232	0.18232
x_3	0.05554	0.05554	1.41615	1.41615
R^2	0.01260	1.00000	0.03600	1.00000

Table 2. Example A: Cross-tabulation of *y* by *x*, and empirical frequency

	Cell-1 $x_1 = 0$ $x_2 = 0$ $x_3 = 0$	**Cell-2** $x_1 = 1$ $x_2 = 0$ $x_3 = 0$	**Cell-3** $x_1 = 0$ $x_2 = 1$ $x_3 = 0$	**Cell-4** $x_1 = 0$ $x_2 = 0$ $x_3 = 1$	**Total**
$y = 0$	102	97	85	99	383
$y = 1$	2	5	2	8	17
Total	104	102	87	107	400
$f(y = 1)$	0.01923	0.04902	0.02299	0.07477	0.04250

As expected, the parameter estimates of linear OLS model (1) and WLS model (5) are the same, but their quality of fit R^2 are drastically different – the first model is of very poor quality while the second has the perfect fit $R^2 = 1$. Similarly, with the logistic regressions: their parameters constructed by the ML (7) or by the linear link (12) coincide as well, but for the logit (7) the quality estimated via the residual and null deviances yields $R^2 = 1 - 135.6/140.65 = 0.036$, while the linear link estimation yields the perfect fit with zero residuals and the coefficient $R^2 = 1$.

Thus, the models (1) and (7) built by all the observations are of a bad quality of fit. The models (5) and (12) are absolutely the same by parameters as (1) and (7), respectively, but being constructed for the frequency of outcome they have the best possible quality of fit. It could be not immediately clear how to interpret such bizarre results.

If the intent is to use the linear model (1) to make predictions, it yields the following values for all 400 observations: 0.01923 if all *x*s equal zero (the intercept of the model (1) in Table 1), 0.04902 if only $x_1 = 1$ and others are zero (the intercept plus the first coefficient of this model in Table 1), 0.02299 if only $x_2 = 1$ and others are zero (the intercept plus the second coefficient of this model), and 0.07477 if only $x_3 = 1$ and others are zero (the intercept plus the third coefficient). The model (5) and logistic models (7) and (12) yield exactly the same prediction values.

Consider the original data in cross-tabulation presented in Table 2. This contingency table shows that there are only four cells of different combinations of the predictors' values. In each cell we see a low incidence rate of the buying intent. The frequency $f(y = 1)$ in total is just 4.25%, and advertising helps to increase it from the benchmark of 1.92% to the maximum of 7.48%.

The predictions by all the models coincide with the original data frequencies in cells. It is so because there are only four different cells, and four parameters in each model, therefore the models not only approximate the data but interpolate via

all points of different combinations of the predictors. It explains why $R^2 = 1$ in the model (5) or (12) built by the four cells data.

The cross-tabulation of the data also explains why the linear OLS and logistic regressions are of such a low quality of fit: because any prediction of the individual value falls in the range of the same frequencies of the event $y = 1$ (lesser even than 8%) shown in Table 2. It means that any model merely cannot produce a substantial (say, about or above 0.5) probability of occurrence of the event $y = 1$. Even summarizing the total impact of all advertising effects (taking the total of all parameters of the model (1) in Table 1) yields just about 12% probability of the event $y = 1$. More adequate for a binary outcome the logistic model (7) with the total impact of all advertising effects (total of all parameters of the model (7) in Table 1) yields probability $p = 1 / (1 + \exp(1.3308)) = 0.2090$, or just about 21% which is also far lower than at least 50% value.

Thus, there are no data to predict occurrence of the event $y = 1$, and it is the reason of so poor R^2 values in Table 1 for the models (1) and (7) built by the individual observations. These models make sense, although there is insufficient data to predict an individual purchase intent. But, in average by all the data in each cell of observations the frequency of the event should be of the values shown in the bottom row of $f\,(y = 1)$ in Table 2.

Example B

The second example is taken from a marketing research problem on the purchasing of a product in households, with the sample size $N = 4{,}175$. The variables are y, binary dependent variable of purchased or not; x_1, binary variable of unaided brand awareness; and x_2, numeric variable of the number of TV spots (from 1 to 4) seen by respondents. Table 3 is arranged similarly to Table 1 and presents results of the regression modeling by this data.

Table 3. Example B: Parameters of the linear, logit, and linear link regressions

	Linear regression model		**Logistic regression model**	
	OLS by observations *a* (1)	**WLS by cells *b* (5)**	**ML by observations *c* (7)**	**Linear link by cells *d* (12)**
Intercept	0.01429	0.01429	−3.80532	−3.76986
x_1	0.07252	0.07252	1.10803	1.05973
x_2	0.01025	0.01025	0.23524	0.22358
R^2	0.01482	0.86970	0.03193	0.96930

Table 4. Example B: Cross-tabulation of *y* and *x*, and predictions by models

	Cell-1 $x_1 = 0$ $x_2 = 1$	**Cell-2** $x_1 = 0$ $x_2 = 2$	**Cell-3** $x_1 = 0$ $x_2 = 3$	**Cell-4** $x_1 = 0$ $x_2 = 4$	**Cell-5** $x_1 = 1$ $x_2 = 1$	**Cell-6** $x_1 = 1$ $x_2 = 2$	**Cell-7** $x_1 = 1$ $x_2 = 3$	**Cell-8** $x_1 = 1$ $x_2 = 4$
$y = 0$	1085	811	830	863	95	89	94	112
$y = 1$	32	29	38	47	6	8	17	19
Total	1117	840	868	910	101	97	111	131
$f\ (y = 1)$	0.02865	0.03452	0.04378	0.05165	0.05941	0.08247	0.15315	0.14504
p linear	0.02453	0.03478	0.04503	0.05527	0.09706	0.10730	0.11755	0.12779
p logit	0.02738	0.03440	0.04312	0.05394	0.07856	0.09737	0.12009	0.14725
p lin.link	0.02802	0.03480	0.04314	0.05338	0.07681	0.09424	0.11513	0.13994

Again, the estimates of parameters in linear OLS model (1) and WLS model (5) are the same, but their quality of fit differs noticeably – the first model is clearly bad with R^2 about 1.5%, while the second is very good with R^2 about 87%. Similarly, with logistic regression estimates: their parameters built in the ML (7) or in the linear link (12) are approximately similar, but the first model has R^2 about 3%, and the second model has R^2 about 97%. So, the models (1) and (7) built by all the observations are of a bad quality of fit. The models (5) and (12) built by the frequency of outcome, are close by parameters to (1) and (7), respectively, but have very high quality of fit.

To explain these features, consider the data cross-tabulation presented in Table 4. There are eight cells of different combinations of the predictor values in this dataset, and each cell shows a low incidence rate of the buying intent with frequency $f\ (y = 1)$. The last three rows in Table 4 present also predictions *p* made by the linear (1) (or (5) with the same results), logit (7), and linear link (12) models.

Predicted values *p* by any model differ from the original frequencies *f* in cells because there are eight observed cells but only three parameters in the regressions (see Table 3), so the models approximate the data. It explains why R^2 for the model (5) or (12) in Table 3 is already not of the perfect fit as it was in Table 1. Still, the linear OLS and logistic regressions are of a poor quality of fit (Table 3) because predictions of the individual values by them are of very low probability *p* of the event $y = 1$ (any of them is less than 15%, see Table 4), which is far below from the middle level of 50%. So, there is no sufficient variability in the data to predict occurrence of the event $y = 1$, and that is the reason of so poor R^2 for the models (1) and (7) built by the individual observations. But, the frequency model (5) and the log-odds of frequency in the linear link model (12) produce probability to belong to each cell and those values *p* are very close to the empirical frequency *f* (Table 4), so R^2 of these models is very high (Table 3).

Table 5. Example B: Correlations of observations and predictions

	y	***f***	***p* linear**	***p* logit**	**p linear link**
y	1.00000	0.13054	0.12173	0.12666	0.12655
f	0.13054	1.00000	0.93255	0.97030	0.96944
p linear	0.12173	0.93255	1.00000	0.98098	0.98261
p logit	0.12666	0.97030	0.98098	1.00000	0.99995
p lin. link	0.12655	0.96944	0.98261	0.99995	1.00000

These data cannot predict individual outcomes, but the probability to belong to each cell is predicted pretty well by any of the considered regressions, as we can see by the values of prediction across the models in each cell of observations in Table 4. The pair correlations of the outcome y with the empirical frequency f and predictions p by all three models are shown in Table 5. We see that all the models yield very similar results, although they cannot reproduce the observed occurrence of the rare event $y = 1$.

Example C

Consider another example taken from marketing research on credit card activity in a bank. There is a dependent binary variable of the card used or not in a time period and sixteen binary predictors describing the card features. The total of $M = 2^{16}$ cells of all combinations of predictor values are possible but actually only $M = 136$ cells were observed for $N = 170$ respondents. Table 6 presents the predictors in the first column and several models in the next numerical columns. The first four models are the regressions (1), (5), (7), and (12) (the same as in Tables 1 and 3). All of them are of a low quality of fit: R^2 in the last row of Table 6 are about 7% for linear (1) and logit (7) models built by all observations, although using frequency in cells shows a slightly better fit with R^2 about 9.5% for the models (5) and (12).

The reason of a low quality of fit even for the models built by the frequency in cells is as follows: in this data there are 136 cells, and 124 of them appeared only once, so the frequencies defined by them are too unsteady. To select more reliable data, take observations of only those cases which correspond to the cells appearing more than once, actually, from 2 to 10 times. Table 6 in the last four columns presents the same four models constructed by the selected data with some variables omitted because they coincide with other withheld predictors in the subsample.

Table 6. Parameters of linear, logit, and linear link regressions, for all data and for cells appeared more than once

	Data by all observations				Data with cells appeared more than once			
	Linear regression		Logistic regression		Linear regression		Logistic regression	
	OLS by observed a (1)	WLS by cells b (5)	ML by observed c (7)	Linear link by cells d (12)	OLS by observed a (1)	WLS by cells b (5)	ML by observed c (7)	Linear link by cells d (12)
Intercept	0.6779	0.6779	0.7899	0.8280	0.7153	0.7083	0.6931	0.9268
Easy	0.1012	0.1012	0.5634	0.4611	−0.0153	−0.0083	0.1542	−0.0795
Simple	−0.0267	−0.0267	−0.1948	−0.1169	0.4389	0.3833	1.3903	2.0024
Frictionless	−0.1616	−0.1616	−0.9097	−0.7575	−0.3819	−0.3750	−1.3863	−1.6199
Protects	−0.0646	−0.0646	−0.3733	−0.3009				
Private	0.1433	0.1433	0.8264	0.6943				
Liability	−0.0713	−0.0713	−0.4138	−0.3402	−0.5729	−0.4792	−19.5661	−2.7483
Relevant	0.0024	0.0024	0.0686	−0.0052	−0.0382	−0.0208	0.4055	−0.1786
Customize	0.1100	0.1100	0.6381	0.5325	0.2813	0.2292	18.3093	1.3466
Personalized	−0.0175	−0.0175	−0.1923	−0.0749				
Compatible	0.0774	0.0774	0.4317	0.3688				
Accepted	0.0907	0.0907	0.5955	0.4425				
Limits	−0.1286	−0.1286	−0.7327	−0.6128	0.2153	0.2083	0.6931	0.9268
Standard	−0.0399	−0.0399	−0.2891	−0.1908				
Control	0.0502	0.0502	0.3326	0.2355	−0.2153	−0.2083	−0.6931	−0.9268
Impressed	0.0432	0.0432	0.2554	0.2112	0.6701	0.5625	37.0241	3.2155
Intrigued	0.0252	0.0252	0.2101	0.0973	−0.4306	−0.3333	−36.3310	−2.1720
R2	0.0721	0.0953	0.0667	0.0941	0.2231	0.9298	0.2403	0.9278

These last models demonstrate improved quality of fit: R^2 are already above 20% for linear (1) and logit (7) models built by the observations and using frequency in cells for the models (5) and (12) raises R^2 to a very high quality of about 93%.

Example D

In another example on customer satisfaction analysis with a credit card, the dependent variable is a binary indicator of problems experienced or not, and all 38 characteristics of the card usage were measured on a 10-point Likert scale and elicited from $N = 604$ respondents. The total number of observed cells is $M = 549$, so the models by the original observations and by cells arranged from 10-point scale values could be of the same quality of fit. Presented in Table 7 are descriptive statistics on this data, with mean values and medians which show that the predictor values are distributed around the level from 7 to 9. Transforming the variables to the binary ones by the criterion of below-above the mean level and finding the cells defined by the binary predictors yields a smaller number of cells $M = 412$, but still it is big enough.

Table 7. Example D: Descriptive statistics for the variables' means and medians

	Mean	Median		Mean	Median
y	0.70	1	x_{20}	7.44	8
x_1	7.91	8	x_{21}	8.22	9
x_2	8.05	9	x_{22}	8.26	9
x_3	7.67	8	x_{23}	8.30	9
x_4	7.83	8	x_{24}	8.26	9
x_5	7.26	8	x_{25}	7.79	8
x_6	7.89	8	x_{26}	7.61	8
x_7	7.38	8	x_{27}	7.43	8
x_8	7.62	8	x_{28}	7.30	8
x_9	7.68	8	x_{29}	7.56	8
x_{10}	7.80	8	x_{30}	7.52	8
x_{11}	7.44	8	x_{31}	8.07	9
x_{12}	7.51	8	x_{32}	8.26	9
x_{13}	6.67	7	x_{33}	8.07	9
x_{14}	7.71	8	x_{34}	6.49	7
x_{15}	7.26	8	x_{35}	7.74	8
x_{16}	7.90	8	x_{36}	7.44	8
x_{17}	8.13	9	x_{37}	7.52	8
x_{18}	7.75	8	x_{38}	8.31	9
x_{19}	7.34	8			

Table 8. Example D: Parameters of linear, logit, and linear link regressions, for all data in 10-point scales, 2-point scales, and for cells appeared more than once

	for 10-point scales				for 2-point scales				for 2-point scales, cells > once			
	linear		logit		linear		logit		linear		logit	
	OLS observ *a* (1)	WLS cells *b* (5)	ML observ *c* (7)	lin.link cells *d* (12)	OLS observ *a* (1)	WLS cells *b* (5)	ML observ *c* (7)	lin.link cells *d* (12)	OLS observ *a* (1)	WLS cells *b* (5)	ML observ *c* (7)	lin.link cells *d* (12)
a_0	0.28	0.28	−1.02	−1.14	0.56	0.56	0.26	0.44	0.52	0.52	0.09	0.14
x_1	−0.03	−0.03	−0.18	−0.16	−0.12	−0.12	−0.72	−0.82	−1.90	−1.84	−38.75	−10.18
x_2	−0.01	−0.01	−0.05	−0.08	0.02	0.02	0.18	0.16				
x_3	0.01	0.01	0.03	0.04	0.07	0.07	0.34	0.44	−0.23	−0.19	−16.34	−2.12
x_4	0.01	0.01	0.04	0.06	0.07	0.07	0.40	0.49	0.48	0.44	17.47	3.22
x_5	0.01	0.01	0.04	0.04	0.06	0.06	0.37	0.37	0.17	0.17	0.82	0.84
x_6	0.02	0.02	0.15	0.10	0.06	0.06	0.33	0.36				
x_7	0.00	0.00	0.01	0.00	0.03	0.03	0.21	0.17				
x_8	−0.02	−0.02	−0.14	−0.11	−0.07	−0.07	−0.41	−0.49				
x_9	−0.02	−0.02	−0.14	−0.13	−0.12	−0.12	−0.73	−0.81	0.48	0.44	17.47	3.22
x_{10}	0.01	0.01	0.07	0.07	0.12	0.12	0.70	0.85				
x_{11}	−0.02	−0.02	−0.10	−0.10	−0.08	−0.08	−0.50	−0.59				
x_{12}	0.01	0.01	0.02	0.04	0.03	0.03	0.19	0.20	0.48	0.44	17.47	3.22
x_{13}	0.00	0.00	−0.01	−0.01	0.00	0.00	−0.03	−0.12	−0.08	−0.08	−0.56	−0.55
x_{14}	−0.02	−0.02	−0.08	−0.09	−0.06	−0.06	−0.30	−0.36				
x_{15}	0.01	0.01	0.07	0.06	0.00	0.00	−0.04	−0.05				
x_{16}	−0.03	−0.03	−0.16	−0.17	−0.15	−0.15	−0.85	−0.97				
x_{17}	0.03	0.03	0.17	0.16	0.08	0.08	0.39	0.52	0.13	0.13	0.57	0.55
x_{18}	0.01	0.01	0.10	0.08	0.00	0.00	0.04	0.03				
x_{19}	0.02	0.02	0.13	0.13	0.12	0.12	0.71	0.79				
x_{20}	−0.02	−0.02	−0.09	−0.10	−0.02	−0.02	−0.18	−0.17				
x_{21}	0.02	0.02	0.11	0.09	0.16	0.16	0.92	1.13	0.48	0.44	17.47	3.22
x_{22}	−0.01	−0.01	−0.04	−0.06	−0.06	−0.06	−0.34	−0.39	0.48	0.44	17.47	3.22
x_{23}	0.05	0.05	0.23	0.25	0.12	0.12	0.60	0.77				

Table 8 (continued).

	for 10-point scales				for 2-point scales				for 2-point scales, cells > once			
	linear		logit		linear		logit		linear		logit	
	OLS observ *a* (1)	WLS cells *b* (5)	ML observ *c* (7)	lin.link cells *d* (12)	OLS observ *a* (1)	WLS cells *b* (5)	ML observ *c* (7)	lin.link cells *d* (12)	OLS observ *a* (1)	WLS cells *b* (5)	ML observ *c* (7)	lin.link cells *d* (12)
x_{24}	0.01	0.01	0.05	0.05	0.06	0.06	0.37	0.38	−0.52	−0.49	−17.66	−3.51
x_{25}	0.00	0.00	−0.02	−0.01	−0.08	−0.08	−0.53	−0.52				
x_{26}	0.00	0.00	−0.01	0.01	0.02	0.02	0.07	0.12				
x_{27}	0.00	0.00	−0.01	0.02	−0.05	−0.05	−0.31	−0.38				
x_{28}	−0.02	−0.02	−0.12	−0.12	−0.10	−0.10	−0.53	−0.69	−0.23	−0.19	−16.34	−2.12
x_{29}	0.01	0.01	0.04	0.06	0.03	0.03	0.15	0.16	0.27	0.27	1.23	1.24
x_{30}	−0.01	−0.01	−0.08	−0.06	0.00	0.00	−0.04	−0.02				
x_{31}	0.00	0.00	−0.01	0.02	−0.04	−0.04	−0.22	−0.27	0.27	0.27	1.23	1.24
x_{32}	0.00	0.00	0.03	0.02	0.11	0.11	0.62	0.76				
x_{33}	0.00	0.00	−0.01	−0.01	−0.05	−0.05	−0.21	−0.33				
x_{34}	0.00	0.00	−0.02	−0.01	−0.02	−0.02	−0.12	−0.24	−0.13	−0.13	−0.97	−0.95
x_{35}	0.01	0.01	0.06	0.07	0.09	0.09	0.55	0.64				
x_{36}	−0.01	−0.01	−0.02	−0.03	−0.01	−0.01	−0.10	−0.05				
x_{37}	0.02	0.02	0.12	0.12	0.03	0.03	0.21	0.20	0.11	0.11	0.54	0.55
x_{38}	0.00	0.00	0.00	0.02	−0.06	−0.06	−0.33	−0.41				
R^2	0.09	0.10	0.08	0.10	0.10	0.14	0.09	0.14	0.13	0.92	0.12	0.87

Presented in Table 8, the first four numerical columns the models built by the predictors in 10-point scale are shown in the same order described for the previous Table 6. The coefficients of multiple determination are shown in the last row, and we see that the quality of all these models is poor, with R^2 not higher than 10%. The models constructed by predictors transformed to the binary 2-point scales are presented in the middle four columns of Table 8. These models are only slightly better by quality of fit. Similarly to the previous example *C*, the low quality of fit even for the models built by the frequency in cells can be explained by the same reason: most of the cells appear only once, so it is hardly possible to predict the dependent variable values in them. To select a more reliable data subset, take observations corresponding to the cells which appeared more than once and there are 210 such cases. In the last four columns Table 8 presents the same four models constructed by the selected data with some variables omitted because of the low variability in the subsample. For the models built by the observations, the quality of fit is still low, about 12-13%. But the quality of fit in both linear and logit models built by the frequencies in cells becomes pretty high, already with the coefficient of multiple determination R^2 about 90%.

It is useful to note that for continuous numerical predictors we can divide their values into several ranges so make the ordinal categorical values and with those to apply approaches described above.

p-value and *D*-value

Related to regression predictions is the problem of insensitive *p*-values in big data. For a sample of several hundred and more observations, practically any test would have a very small *p*-value which indicates a significant difference in the compared values (Goodman, 2008; Berdie, 2012; Robertson & Kaptein, 2016; Wasserstein & Lazar, 2016; Johnson et al., 2017). In Demidenko (2016), a regular *p*-value is criticized and another criterion, namely *D*-value, is proposed as an alternative for meaningful hypotheses testing which can yield reasonable results for practical applications. For instance, checking hypotheses on the difference of the sample mean $\bar{x}$ or an individual observation x_i from the population mean μ with unknown variance σ^2 standard deviation σ can be done by the following *t*-tests, respectively:

$$\mathrm{t}\left(\bar{x}\right)=\frac{\bar{x}-\mu}{\sqrt{\sigma^2/N}}=\frac{\bar{x}-\mu}{\sigma}\sqrt{N},\quad \mathrm{t}\left(x_i\right)=\frac{x_i-\mu}{\sigma}. \tag{13}$$

Checking for the mean $\bar{x}$ depends on the sample size proportionally to $\sqrt{N}$, and the value t($\bar{x}$) can grow with large N, so regardless the difference $\bar{x}-\mu$ the p-value can be very small producing an impression that whatever difference is statistically significant. Demidenko (2016) essentially suggests instead of t($\bar{x}$) to use t(x_i) without the term $\sqrt{N}$, as checking for an outlier by an individual observation x_i.

The idea to use the standard deviation σ instead of the standard error ($se=\sigma/\sqrt{N}$) can be seen in the context of confidence intervals for prediction versus fitting in simple linear regression. As it is well-known (for instance, Weisberg, 1985, p. 22, p. 281), the variances of a fitted value $\hat{y}_i$ and of a predicted value $\hat{y}_*$ equal, respectively

$$\operatorname{var}\left(\hat{y}_i\right)=\sigma^2\left(\frac{1}{N}+\frac{\left(x_i-\bar{x}\right)^2}{S_x^2}\right),\quad \operatorname{var}\left(\hat{y}_*\right)=\sigma^2\left(1+\frac{1}{N}+\frac{\left(x_*-\bar{x}\right)^2}{S_x^2}\right), \qquad (14)$$

where σ^2 is residual variance, and $\bar{x}$ and S_x^2 are the predictor's mean and centered sum of squares, respectively. The variances (14) differ by one σ^2 added for a new observation at x_* used in place of an observed x_i. For the fit and prediction at the point of mean value $\bar{x}$ the last items in each of the formulae (14) equal zero, so taking square root of the variances yields the standard errors for the fit and prediction, respectively

$$\operatorname{se.fit}\left(\hat{y}_i\right)=\sigma\frac{1}{\sqrt{N}},\quad \operatorname{se.pred}\left(\hat{y}_*\right)=\sigma\sqrt{1+\frac{1}{N}}\approx\sigma, \qquad (15)$$

where the last approximation works even for a relatively medium sample size N. The first formula in (15) defines the regular standard error used for hypotheses testing t($\bar{x}$) in (13), which is producing p-values. The second formula in (15) equals the standard deviation, so it corresponds to the second formula in (13) which is yielding the D-values suggested in Demidenko (2016) for hypotheses checking. Thus, D-value can be interpreted in terms of a p-value rather for a predicted than fitted estimation with the corresponded hypotheses testing.

Summary

If a model is built by individual observations in a given dataset it could show a poor quality of fit and bad predictions of individual observations; but if to use frequencies of the outcome and possible combinations of the predictors, we build the same model with the same parameters, however, with a high quality of fit and precise predictions. The reasons for such results correspond to the definition of regression as the expectation of the outcome *y* subject to the given predictors' values, E(*y* | *x*). So, for each unique combination of the independent variables *x* values the regression predicts the average of the outcome *y*, or its frequency in the range of the cells' combinations. Adequate interpretation is important for understanding of the models' behavior, especially for prediction of the rare events. Linear and logistical regressions models are used for illustration of the results. The considered problems are also completed with the explanation on the *p*-value and *D*-value in relations to the predictions by regression models. The ideas of collapsing individual observations into cells could be analyzed in data mining. Meaningful usage of regressions is absolutely important for practical needs, and the obtained results help to a better understanding of properties of multiple regression, are valuable for theoretical consideration and practical applications of regression modeling, analysis, and prediction.

References

Andersen, P. K., & Skovgaard, L. T. (2010). *Regression with linear predictors*. New York: Springer. doi: 10.1007/978-1-4419-7170-8

Berdie, D. (2012, January). Data use: Significant differences. *Quirk's media magazine*, article ID 20120104. Available from https://www.quirks.com/articles/data-use-significant-differences

Demidenko, E. (2016). The *p*-value you can't buy. *The American Statistician*, *70*(1), 33-38. doi: https://doi.org/10.1080/00031305.2015.1069760

Goodman, S. (2008). A dirty dozen: Twelve *p*-value misconceptions. *Seminars in Hematology*, *45*(3), 135-140. doi: 10.1053/j.seminhematol.2008.04.003

Grafarend, E. W., & Awange, J. L. (2012). *Applications of linear and nonlinear models: Fixed effects, random effects, and total least squares*. New York: Springer. doi: 10.1007/978-3-642-22241-2

Härdle, W. K., & Simar, L. (2012). *Applied multivariate statistical analysis.* New York: Springer. doi: 10.1007/978-3-662-45171-7

Hilbe, J. M., & Robinson, A. P. (2013). *Methods of statistical model estimation*. Boca Raton, FL: Chapman and Hall/CRC Press.

Izenman, A. J. (2008). *Modern multivariate statistical techniques*. New York: Springer. doi: 10.1007/978-0-387-78189-1

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association, 112*(517), 1-10. doi: 10.1080/01621459.2016.1240079

Kendall, M. G., & Stuart, A. (1973). *The advanced theory of statistics* (3rd edition, Vol. 2). New York: Hafner.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer. doi: 10.1007/978-1-4614-6849-3

Lipovetsky, S. (2012). Interpretation of Shapley value regression coefficients as approximation for coefficients derived by elasticity criterion. *Proceedings of the 2012 Joint Statistical Meeting of the American Statistical Association,* 3302-3307, San Diego, CA.

Lipovetsky, S. (2013). How good is best? Multivariate case of Ehrenberg-Weisberg analysis of residual errors in competing regressions. *Journal of Modern Applied Statistical Methods, 12*(2), 242-255. doi: 10.22237/jmasm/1383279180

Lipovetsky, S. (2015). Analytical closed-form solution for binary logit regression by categorical predictors. *Journal of Applied Statistics, 42*(1), 37-49. doi: 10.1080/02664763.2014.932760

Lipovetsky, S. (2018). MaxDiff choice probability estimations on aggregate and individual level. *International Journal of Business Analytics, 5*(1), 55-69. doi: 10.4018/IJBAN.2018010104

Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry, 17*(4), 319-330. doi: 10.1002/asmb.446

Lipovetsky, S., & Conklin, M. (2010). Meaningful regression analysis in adjusted coefficients Shapley value model. *Model Assisted Statistics and Applications, 5*(4), 251-264. doi: 10.3233/MAS-2010-0170

Lipovetsky, S., & Conklin, M. (2014). Best-worst scaling in analytical closed-form solution. *Journal of Choice Modelling, 10,* 60-68. doi: 10.1016/j.jocm.2014.02.001

Lipovetsky, S., & Conklin, M. (2015). Predictor relative importance and matching regression parameters. *Journal of Applied Statistics*, *42*(5), 1017-1031. doi: 10.1080/02664763.2014.994480

McCullagh, P., & Nelder, J. A. (1999). *Generalized linear models* (2nd edition). New York: Chapman & Hall.

Robertson, J., & Kaptein, M. (Eds.) (2016). *Modern statistical methods for HCI*. Cham, Switzerland: Springer. doi: 10.1007/978-3-319-26633-6

Train, K. (2003). *Discrete choice methods with simulation*. New York: Cambridge University Press.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129-133. doi: 10.1080/00031305.2016.1154108

Weisberg, S. (1985). *Applied linear regression* (2nd edition). New York: Wiley.

Wilson, J. R., & Lorenz, K. A. (2015). *Modeling binary correlated responses using SAS, SPSS and R*. New York: Springer. doi: 10.1007/978-3-319-23805-0