Direct ascription of missing values in survey research data

Mentor:
Martin Dimov

Working team:

Assen Tchorbajieff, Vasil Kolev, VeskaNoncheva, Venelin Valkov, Dimitar Fidanov, Elica Ilieva, Maria Dobreva, Maroussia Bojkova

Introduction

- Company’s overview
- Definition of the problem
- Task description
- Materials provided

So far we've done…

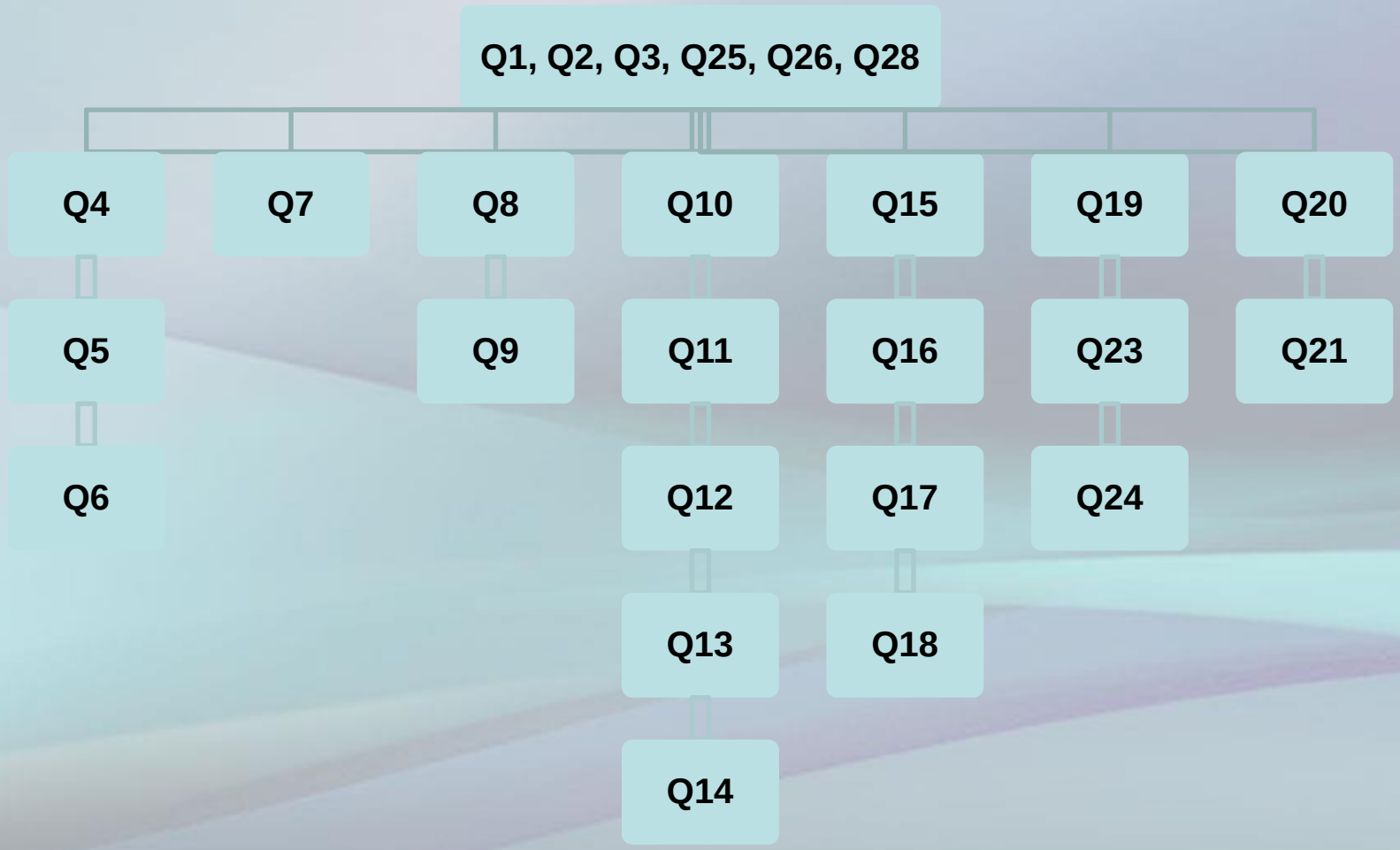
- Logical data separation
- Data predictions
- Logical scheme
- Calculated the frequency
- Made software program inR
- Correspondence analysis
- Random forest

Data separation

What are we looking for?

- Logical connections according to the request list (Graphical scheme)
- Indentifying the problem(Table)
- Identification of the metadata
- Data / predictors(Software)
- Searching for regression statistical relations

Relations scheme



Identifying the problem

Question 9	Frequency	Question 11	Frequency
9.1	31, 01 %	11.1	35, 2 %
9.2	24, 69 %	11.2	50, 3 %
9.3	*8, 91 %*	11.3	51, 4 %
9.4	38, 47 %	11.4	52, 9 %
9.5	35, 49 %	11.5	21 %
9.6	43, 17 %	11.6	9, 2 %
9.7	51, 39 %	11.7	12 %
9.8	42, 47 %	11.8	48, 6 %
9.9	33, 33 %	11.9	50, 4 %
9.10	18, 29 %	11.10	25, 1 %
9.11	21, 43 %	11.12	32, 7 %
9.12	16, 68 %	11.13	48, 7 %
9.13	43,17 %		52 %
9.14	35 %		
9.15	44, 26 %		
9.16	39, 8 %		
9.17	35, 16 %		

Software program in R

- Compile files according logical scheme
- Removes rows with NA/0
- Select pre-requested variables
- Sorted by columns
- Being able to use for multiple computations

First step failure

We offered very simple model for preliminary session:

- Selection by sex (Male/female)
- Predictors
- Age
- What's your average spent for food and grocery
- Fraction of income/members of the family
- How the price affects on clients choice

Model failure (Females)

Residuals:

Min	1Q	Median	3Q	Max
-2.9598	-1.2541	-0.4466	0.7706	5.2061

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	6.502e-01	9.894e-01	0.657	0.5128	
Age	5.477e-02	2.558e-02	2.141	0.0351	*
IncimeFrac	1.502e-04	8.755e-05	1.716	0.0897	.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.779 on 87 degrees of freedom
Multiple R-squared: 0.09073, Adjusted R-squared: 0.06982
F-statistic: 4.34 on 2 and 87 DF, p-value: 0.01597

Reasons

- Software defect due to Croatian language encoding

The improved file descriptions are:

- Size of dataset:
  - Male – 85 results (last 50)
  - Female – 91 results (last 48)
- Size of available for prediction values:
  - Male – 94 results (last 5)
  - Female – 103 results (last 4)

Running the data

- The model works with the assumption of normality and log-normality.
- The “Age” predictor proves useless
- We cannot confirm the “male” data due to strong saving habits

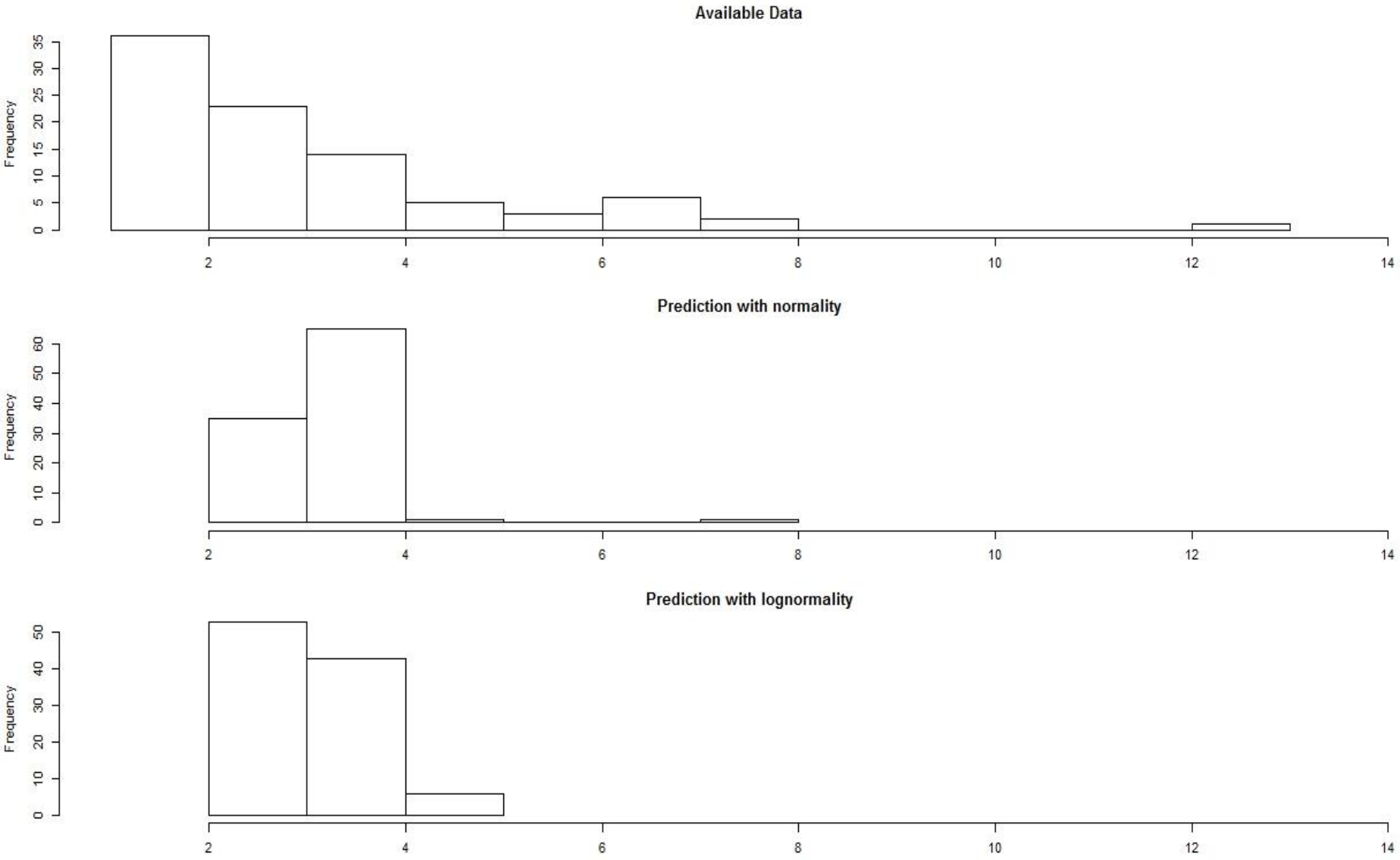
The prediction

- The possible missing values are restored by reversing formula for I-thelement:

X(i) = round(intercept + sum(pred_coeff(i)*pred(i)))

- The values are verified by the rest of the available data in the row because of avoidance requirements of the repeat
- It's excluded by logical assumption:
- Normal distribution - 39
- Log-normal distribution 34

Result statistics



Improvements

- Identifying the 'weak' data
- We discovered a large outlier:
- 1 record of low income vs. lack of saving habits
- The record is removed
- The regression is running again

Results

Deviance Residuals:

Min	**1Q**	**Median**	**3Q**	**Max**
-2.38964	-0.94833	-0.08752	0.75072	3.12686

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	
(Intercept)	3.7106	1.7826	2.082	0.04050	*
Spend	0.6053	0.2231	2.713	0.00813	**
IncomeFrac	-0.6290	0.1517	-4.146	8.21e-05	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter forgaussianfamily taken to be 1.648235)

Null deviance: 168.75 on 84 degrees of freedom

Residual deviance: 135.16 on 82 degrees of freedom

AIC: 288.64

Dependency only on Income Fraction

Residuals:

Min	1Q	Median	3Q	Max
-2.4476	-1.4110	-0.1183	0.7719	5.2474

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.679e+00	2.897e-01	9.248	1.27e-14	***
IncomeFrac	1.756e-04	8.849e-05	1.984	0.0504	.

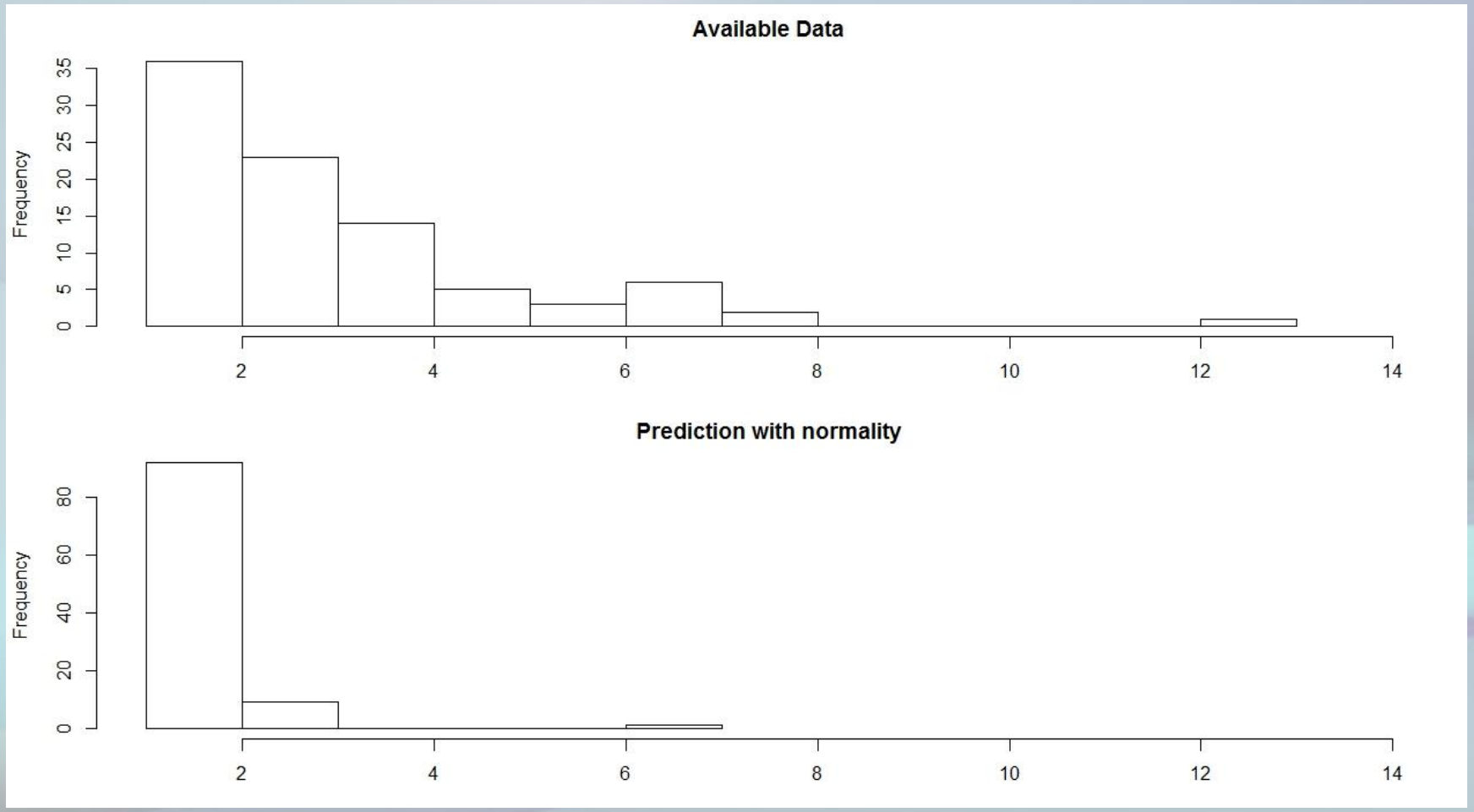
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.815 on 88 degrees of freedom

Multiple R-squared: 0.04282, Adjusted R-squared: 0.03194

F-statistic: 3.937 on 1 and 88 DF, p-value: 0.05036

Distributions



To do next:

- Generate predictions for all columns and subdata
- Check control logic for coincidences
- If no data left rethink the predictors
- Else input complete data rows and
  rerun the model again.
- Repeat recursively until stops.

Missing value prediction with correspondence analysis

- Categorical analysis
- Categorical data/variables
- Communicate complex tables
- Easy 2D/3D plotting
- Similar users – similar behavior (CF)

Thank you for the attention!