Multidimensional project

PETS FOOD LABEL USE DATA FDA

Eugeniah Arthur

Bowling Green State University

June 2020

Table of Contents

https://www.overleaf.com/project/5ef5f24eca899500013e8530

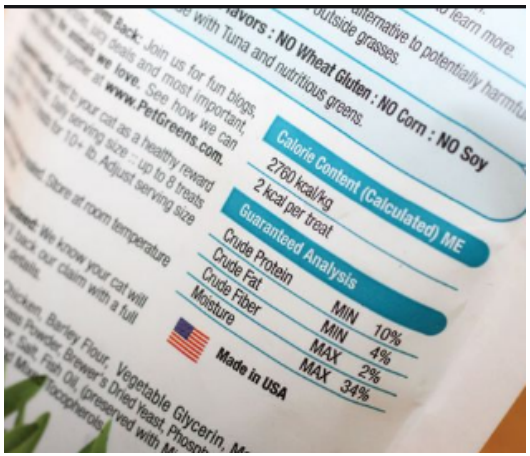
Introduction

Pet food has become one of the leading food products that has become of very great essence to this generation. Increasing research has shown that there is a need for pet owners to take note of the food they buy for their pets. This has left a question of what purchasers really consider from the label the first time they purchase a particular pet food. This study seeks to investigate this.

Background Of data

The data was part of the 2008 Health and Diet Survey Food and Drug Administration data on Pet food Label Use . It was based on a random - digit-dialing telephone survey of 2584 non-institutionalized adults in all 50 states of the united states with the district of Columbia included.
Item K6 data is what is of essence in this study . It deals with what first time buyers of a particular pet food consider when they look at the product label. They were asked whether they often, sometimes, rarely, never use the food label for these purposes. It is noticed that some people chose not to answer to this item questions and hence there will be a need for data cleaning.

image of pet food labels



Item K6

The questions in the item are:

- K6A- To figure out how much of the pet food you should feed your dog/dogs/cat/cats/dog and cat/dogs and cat/dog and cats/dogs and cats at a single serving?
- K6B- To see how high or low the pet food is in things like calories, protein, fat, etc.?
- K6C- To determine if a product meets your pets' nutritional needs?
- K6D -To compare different pet food items with each other?
- K6E - To see if something said in advertising or on the package is actually true?
- K6F - To see if there is an ingredient that your pet should avoid?

These are the broad variables that the FDA considers may be good measures that a pet owner may consider when buying a particular pet food product for the first time. .

Objectives of the Study

The main objective of the study is to :

- Harness the majority of information that influences the use of pet food products labels by the use of principal component analysis.
- Identify the factors that influence the use of pet food product label using factor analysis

Principal Component Analysis

The main idea of principal component analysis is to reduce the dimension of the variables to a small dimension that will account for a majority of the variation in the data. Creating new components using a linear combination of the original data. It does not assume any dimension beforehand to reduce it to. The principal components is a linear combination of the original variables based on average weights. Mathematically,

$$Z_i = W_i Y_I$$

Where Z_i is the new components and W_i is the weight given to each original variable Y_i.

Factor Analysis

Factor Analysis also reduces the dimension of the variables based on some latent variables. These latent variable cannot be directly measured, they are assumed to be what is influencing the variation in the data. Factor analysis identifies the latent variables that may be causing the variation in original variables. Thus, it is a linear combination of the latent variables. Mathematically,

$$X = QF + \mu + U$$

- X will be the value of the observed variables
- Q is the weight of the factor on each the variable
- F will be the latent factors
- U Unique factors
- μ is the mean of each variable

Assumptions of Factor Analysis

- The variables are linearly related to the factors and errors
- Observations are independent
- The model assumes that the variables are determined by only common factors and unique factors . The unique factors are uncorrelated to themselves and the common factors.
- $E(F) = 0$ and $Cov(F) = I, Cov(FU) = 0, Cov(U) = \lambda$

Confirmatoy Factor Analysis

Typically , exploratory factor analysis just gives you an idea of the number of factors that may be present in the data. However, CFA uses statistical technique to come to a more satisfying conclusion of whether those number of factors are the actual factors and the relationship between these factors and the variables.

Cursory look at the data percentages

		OFTEN	SOMETIMES	RARELY	NEVER	DK/NS	REF
a.	to figure out how much of the pet food you should feed your {dog/dogs/cat/cats/dog and cat/dogs and cat/dog and cats/dogs and cats} at a single serving?	30%	15%	17%	38%	*	*
b.	to see how high or low the pet food is in things like calories, protein, fat, etc.?	31	18	16	35	*	0
c.	to determine if a product meets your pets' nutritional needs?	37	26	10	26	1	0
d.	to compare different pet food items with each other?	38	18	12	31	*	0
e.	to see if something said in advertising or on the package is actually true?	22	22	16	39	1	0
f.	to see if there is an ingredient that your pet should avoid?	35	14	11	39	1	0

Figure: Summary from EDA Report!

Data Cleaning

All missing values in the data were removed using a SAS code. Only 1005 data set were left for further analysis

The FACTOR Procedure

Input Data Type	Raw Data
Number of Records Read	1005
Number of Records Used	1005
N for Significance Tests	1005

Figure: Data Retained

Summary of data

Ã¯..K6A	K6B	K6C	K6D
Min. :1.000	Min. :1.000	Min. :1.000	Min. :1.000
1st Qu.:1.000	1st Qu.:1.000	1st Qu.:1.000	1st Qu.:1.000
Median :3.000	Median :3.000	Median :2.000	Median :2.000
Mean :2.612	Mean :2.574	Mean :2.271	Mean :2.459
3rd Qu.:4.000	3rd Qu.:4.000	3rd Qu.:4.000	3rd Qu.:4.000
Max. :4.000	Max. :4.000	Max. :4.000	Max. :4.000

K6E	K6F
Min. :1.000	Min. :1.000
1st Qu.:2.000	1st Qu.:1.000
Median :3.000	Median :3.000
Mean :2.802	Mean :2.664
3rd Qu.:4.000	3rd Qu.:4.000
Max. :4.000	Max. :4.000

Figure: Five number summary!!

Explained Variance

```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     1.9091 0.8454 0.72496 0.66493 0.61586 0.54198
## Proportion of Variance 0.6074 0.1191 0.08759 0.07369 0.06321 0.04896
## Cumulative Proportion  0.6074 0.7266 0.81414 0.88783 0.95104 1.00000
```

Figure: Variance Explained

The first two PCS would contain 72.66% of the variation in the data. The remaining PCs account for very little variations in the data.

Plots

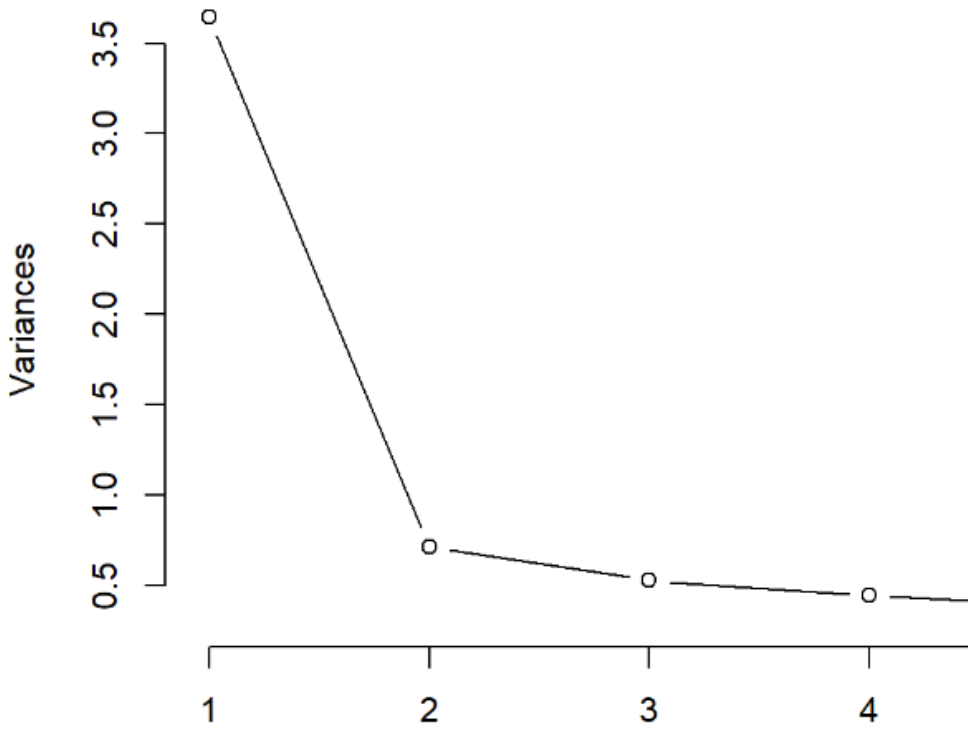


Figure: Screeplot

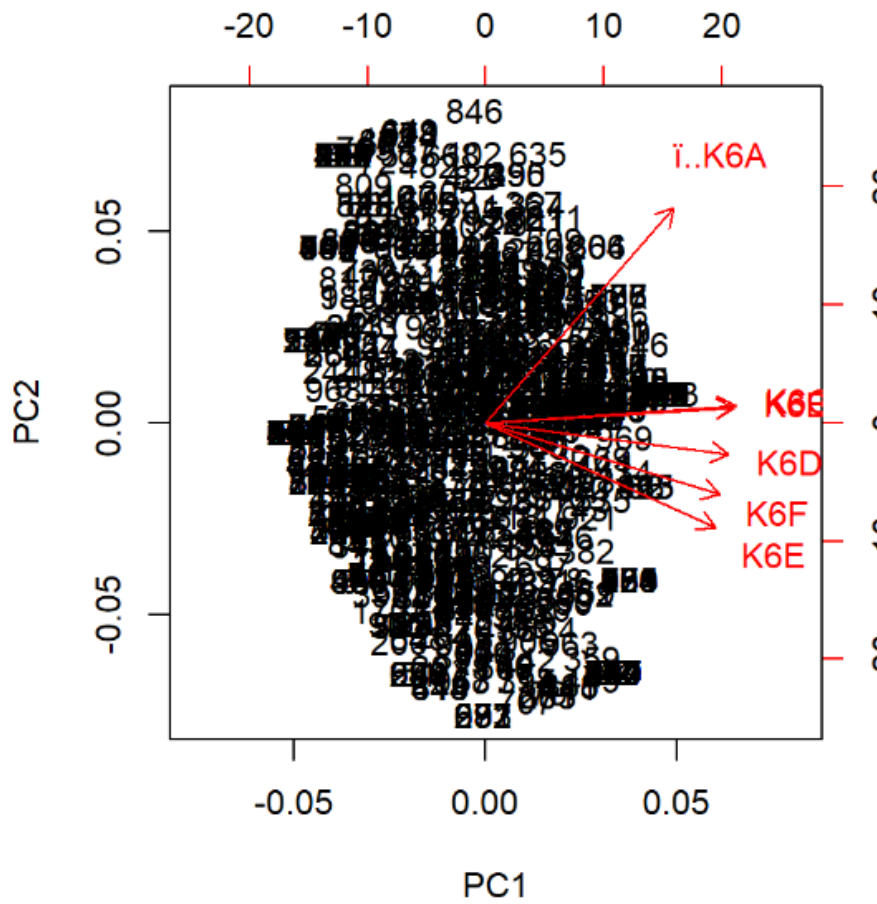


Figure: biplot

Summary from PCA

- The scree plot shows that the majority of the variation in the data is accounted fro by the first two principal components. This is about 72% of the variation in the information gained from the survey of the use of the pet food product label .
- K6A loads fairly equally and positively on both PC1 and PC2 . That component does not seem to be correlated to other variables. However, K6C and K6B have the same influence on the two components. However, it is seen that , these two influence PC1 more and have very little influence on PC2 . K6D, K6F and K6E all have a negative influence on PC2 but has a positive influence on PC1. All the variables except K6A seems to be correlated with each other and have about the same influence on the PC2.

Results from EFA Principal Axis Method

Eigenvalues of the Reduced Correlation Matrix: Total = 2.86636856 Average = 0.4777280

	Eigenvalue	Difference	Proportion	Cumulativ
1	3.14231971	3.03694992	1.0963	1.096
2	0.10536979	0.15637818	0.0368	1.133
3	-.05100839	0.00673845	-0.0178	1.115
4	-.05774684	0.05421792	-0.0201	1.095
5	-.11196476	0.04863618	-0.0391	1.056
6	-.16060094		-0.0560	1.000

2 factors will be retained by the NFACTOR criterion.

Figure: Proportion Explained

Variance Explained by Each Factor

Factor1	Factor2
3.1423197	0.1053698

Final Communality Estimates: Total = 3.247689

K6A	K6B	K6C	K6D	K6E	K6F
0.30882526	0.63521677	0.65179961	0.57552445	0.54008561	0.53623780

Figure: Final Communality Estimate

Factor Pattern - Principal Axis Method

Factor Pattern				
	Factor1		Factor2	
K6A	54	*	12	
K6B	79	*	12	
K6C	80	*	13	
K6D	76	*	-4	
K6E	71	*	-19	
K6F	72	*	-13	
Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.47 are flagged by an '*'.				

Rotated Factor Pattern				
	Factor1		Factor2	
K6A	48	*	29	
K6B	65	*	46	
K6C	67	*	46	
K6D	52	*	56	*
K6E	37		63	*
K6F	42		60	*
Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.47 are flagged by an '*'.				

Variance Explained by Each Factor	
Factor1	Factor2
1.6772870	1.5704024

Figure: Unrotated Factor Pattern

Results from EFA Maximum LIkelihood Method

Significance Tests Based on 1005 Observations			
Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors	15	2699.8350	<.0001
HA: At least one common factor			
H0: 2 Factors are sufficient	4	3.9166	0.4174
HA: More factors are needed			

Figure: Significance test of two factors

Final Communality Estimates and Variable Weights		
Total Communality: Weighted = 11.702024 Unweighted = 3.597837		
Variable	Communality	Weight
K6A	0.30533448	1.43953016
K6B	0.66411748	2.97725293
K6C	0.73044441	3.70980729
K6D	0.57031713	2.32729942
K6E	0.80613912	5.15833387
K6F	0.52148454	2.08979980

Figure: Final Communality Estimate

Factor Pattern - ML Method

Factor Pattern				
	Factor1		Factor2	
K6A	52	*	-18	
K6B	78	*	-23	
K6C	80	*	-31	
K6D	75	*	-6	
K6E	81	*	40	
K6F	72	*	3	
Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.47 are flagged by an '*'.				

Figure: Unrotated Factor Pattern

Rotated Factor Pattern				
	Factor1		Factor2	
K6A	51	*	22	
K6B	73	*	37	
K6C	79	*	32	
K6D	59	*	48	*
K6E	32		84	*
K6F	50	*	52	*
Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.47 are flagged by an '*'.				

Variance Explained by Each Factor		
Factor	Weighted	Unweighted
Factor1	6.13276867	2.11451866
Factor2	5.56925501	1.48331850

Summary from Exploratory Factor Analysis

- The exploratory factor analysis seems to show that two factors are influencing the variability in the data. Thes will tentatively be termed content factor and Economic factor
- The unrotated factor pattern shows that just one factor may be important. This is not surprising because the questions were designed to account for how pet food product labels are used.
- Both methods shows that K6A,K6B and K6C are strongly influenced by the Content factor whereas k6E is strongly influenced by Economic factor. K6D seems to be influenced by both factors very highly . The ML method shows K6F is strongly influenced by both factors as well.

Model for Confirmatory Factor Analysis

Initial Estimates for Linear Equations									
K6A	=	LV1F1	(.)	F1	+		1	E1	
K6B	=	LV2F1	(.)	F1	+		1	E2	
K6C	=	LV3F1	(.)	F1	+		1	E3	
K6D	=	LV4F1	(.)	F1	+	LV4F2 (.) F2 +		1	E4
K6E	=	LV5F2	(.)	F2	+		1	E5	
K6F	=	LV6F2	(.)	F2	+		1	E6	

Figure: Equations for Prin Method

Initial Estimates for Linear Equations									
K6A	=	LV1F1	(.)	F1	+		1	E1	
K6B	=	LV2F1	(.)	F1	+		1	E2	
K6C	=	LV3F1	(.)	F1	+		1	E3	
K6D	=	LV4F1	(.)	F1	+	LV4F2 (.) F2 +		1	E4
K6E	=	LV5F2	(.)	F2	+		1	E5	
K6F	=	LV6F1	(.)	F1	+	LV6F2 (.) F2 +		1	E6

Figure: Equation for ML

Model Assessment Index

Index	Estimate Prin	Estimate ML	Threshold
Power	0.774	0.718	0.8
RMSEA	0.0391(0.0167, 0.0622)	0.0152 (0.00,0.0457)	0.055
SRMR	0.0124	0.0073	0.055
CFI	0.9960	0.995	0.94

Table: Indexes

The power of test seems to be relatively low. However, the indexes shows that the model is a good fit and its estimates may be adequate.

Standardized Equation Estimates-Prin

Standardized Results for Linear Equations						
K6A =	0.5533 (**)	F1 +	1.0000	E1		
K6B =	0.8274 (**)	F1 +	1.0000	E2		
K6C =	0.8402 (**)	F1 +	1.0000	E3		
K6D =	0.3801 (**)	F1 +	0.4119 (**)	F2 +	1.0000	E4
K6E =	0.7689 (**)	F2 +	1.0000	E5		
K6F =	0.7721 (**)	F2 +	1.0000	E6		

Standardized Effects in Linear Equations

Variable	Predictor	Parameter	Estimate	Standard Error	t Value	Pr > |t|
K6A	F1	LV1F1	0.55328	0.02434	22.7309	<.0001
K6B	F1	LV2F1	0.82740	0.01393	59.3981	<.0001
K6C	F1	LV3F1	0.84025	0.01352	62.1604	<.0001
K6D	F1	LV4F1	0.38014	0.06988	5.4395	<.0001
K6D	F2	LV4F2	0.41188	0.07116	5.7885	<.0001
K6E	F2	LV5F2	0.76887	0.01814	42.3851	<.0001
K6F	F2	LV6F2	0.77207	0.01807	42.7330	<.0001

Standardized Equation Estimates-ML

Standardized Effects in Linear Equations						
Variable	**Predictor**	**Parameter**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
K6A	F1	LV1F1	0.55384	0.02431	22.7821	<.0001
K6B	F1	LV2F1	0.82590	0.01395	59.2043	<.0001
K6C	F1	LV3F1	0.84145	0.01345	62.5698	<.0001
K6D	F1	LV4F1	0.51693	0.07514	6.8794	<.0001
K6D	F2	LV4F2	0.29404	0.07639	3.8493	0.0001
K6E	F2	LV5F2	0.87828	0.06305	13.9295	<.0001
K6F	F1	LV6F1	0.38196	0.09870	3.8699	0.0001
K6F	F2	LV6F2	0.40096	0.09998	4.0103	<.0001

Standardized Results for Exogenous Variables

Standardized Results for Variances of Exogenous Variables						
Variable Type	**Variable**	**Parameter**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Latent	**F1**		1.00000			
	F2		1.00000			
Error	**E1**	VARE1	0.69388	0.02693	25.7618	<.0001
	E2	VARE2	0.31541	0.02305	13.6831	<.0001
	E3	VARE3	0.29398	0.02272	12.9416	<.0001
	E4	VARE4	0.42102	0.02404	17.5164	<.0001
	E5	VARE5	0.40884	0.02789	14.6566	<.0001
	E6	VARE6	0.40391	0.02790	14.4779	<.0001

Figure: Prin Method

Standardized Results for Variances of Exogenous Variables						
Variable Type	**Variable**	**Parameter**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Latent	**F1**		1.00000			
	F2		1.00000			
Error	**E1**	VARE1	0.69326	0.02693	25.7450	<.0001
	E2	VARE2	0.31790	0.02304	13.7962	<.0001
	E3	VARE3	0.29195	0.02263	12.8999	<.0001
	E4	VARE4	0.42751	0.02463	17.3559	<.0001
	E5	VARE5	0.22863	0.11075	2.0643	0.0390
	E6	VARE6	0.47286	0.03051	15.4999	<.0001

Figure: ML

Summary from Standardized results

- The T-Values of the standardized estimates are all significant.
- K6B ad K6C are most influenced by the Content Factor,their standardized estimates are very high and K6E is highly influenced by Economic factors
- The estimates K6D and K6F shows that there is a high influence of both factors.
- There are no negative estimates. The estimates can be interpreted as for every additional increase in the factor, the variable increases on average by the estimate value while holding all other variables constant.
- The influence of the unique factor is significant. That is each of the factors have individual influences that affect them.

Reasons for discrepancy in results of the Principal Axis Method and the ML method

- The Maximum Likelihood Method assumes a multivariate normal distribution however the Principal axis method does not have that assumption and so when your data stongly violates this assumption then the Principal axis method is a better option
- The ML Method provides indexes of the goodness of fit of the model . It also permits statistical significance testing of the number of factor , the factor loading and correlations among factors and the computation of confidence intervals.

Conclusion

- 72% of the information on use of pet product food label use can be found from the first two principal components found.
- The two factors can be summarized as the Content factor and Economic factors. .
- The amount of food per single serving (K6A) , how high or low pet food is in nutrients (K6B) , whether product meets nutritional needs(K6C) loads very highly on the Content factor while Advertising specification (K6E) is influenced by the Economic factor.
- Comparing different pet foods loads and ingredient to be avoided are influenced highly by both the Economic and Content factor .This results is not surprising

References

- O'Rourke, Norm, and Larry Hatcher. 2013. A Step-by-Step Approach to Using SAS® for Factor Analysis and Structural Equation Modeling, Second Edition. Cary, NC: SAS Institute Inc.
- FDA, 2013 Health and Diet Survey.


Thank you!