The Food Consumption Analysis

BACKGROUND FOR STUDY

The CHIS 2005 Adult Survey contains data on the individual records from the adult component of 2005 California Health Interview Survey. That is the population based random digital dial telephone survey. It collects the extensive information for all age groups on health status, health conditions, health related behaviors, health insurance coverage, access to health care services and other health related issues. Within each household separate interviews were conducted with a randomly selected adult (age 18 and over), adolescents ages (12-17), and parents of children of ages (0-11). A printable version of the adult questionnaire can be found on the CHIS web site at http://www.chis.ucla.edu/topics.html.

ABSTRACT

- The habitual food consumption of the individuals has very important consequences for health of those individuals. Based on the CHIS 2005 Adult Survey with the sample of 16 variables and 42356 cases, the analysis of the food consumption on the weekly basis showed the high correlation between the body measures: height and weight and the consumption of sugar, fruit and vegetables. Canonical Correlation shows the highest correlation of 0.6155 between the height and weight and the consumption of foods: fruit, salad, fries, potatoes, beans, vegetables, sugar, soda, juice, flavored drinks, cake/pie/cookies and ice cream/frozen desserts. The increasing relationship exists between height and weight and the consumption of sugar, fries and potatoes. The negative relationship exists between the height and weight and the consumption of salad, beans, fruit and vegetables. Factor Analysis identified four independent factors explaining 63% of the total variance: factor 1 Fruits, Vegetables, factor 2 Sugar in Drinks, factor 3 Cakes and Deserts and factor 4 Potatoes and Beans. Discriminant Analysis classified respondents into two groups: overweight and not with accuracy of 93.34%.The results were confirmed by Logistic Regression with addition of age. Only height and weight were identified as predictors of being overweight or not with accuracy of 99.43%. Manova shows the statistical difference between groups in age and BMI when considering the effect of health condition and exercising.
- **Conclusions**: The statistical analysis showed the overall effect of the consumed food on body measures: height and weight.

VARIABLES USED

There are the following matrix (continuous) variables

- FV-daily servings of fruits and vegetables
- FVNB-daily servings of fruits and vegetables except beans
- FVNF-daily servings of fruits and vegetables except French fries
- FVNFB-daily servings of fruits and vegetables except French fries and beans
- SUG-teaspoons of added sugar consumed per day
- HGHTI_P-height in inches
- WGHTP_P-weight in lbs
- BMI_P-body mass index
- AE_FRUIT-number of times # eating fruit per week
- AE_SALAD-number of times # eating salad per week
- AE_FRIES-number of times # eating fries per week
- AE_POTAT-number of times # eating potatoes per week
- AE_BEANS-number of times # eating beans per week
- AE_VEGI-number of times # eating vegetables per week
- AE_SODA-number of times # drinking soda per week
- AE_JUICE-number of times # drinking 100% fruit juices per week
- AE_FLAV-number of times # drinking fruit-flavored drinks per week
- AE_CAKE-number of times # eating cake/pie/cookies per week
- AE_FROZ-number of times # eating ice cream/frozen desserts per week

DESCRIPTIVE STATISTICS

Variable	*N*	*Mean*	*Std Dev*	*Median*	*Minimum*	*Maximum*
FV	42356	5.22	1.74	4.98	0.67	19.13
FVNB	42356	4.82	1.69	4.59	0.66	18.76
FVNF	42356	5.04	1.71	4.82	0.67	18.65
FVNFB	42356	4.79	1.64	4.58	0.66	17.85
SUG	42356	14.58	7.27	12.98	3.48	73.22
HGHTI_P	42356	66.33	4.27	66.00	42.00	77.00
WGHTP_P	42356	166.13	39.78	160.00	90.00	330.00
BMI_P	42356	26.46	5.58	25.61	11.79	99.34
AE_FRUIT	42356	8.01	6.98	7.00	0.00	84.00
AE_SALAD	42356	5.00	4.02	5.00	0.00	35.00
AE_FRIES	42356	0.69	1.31	0.00	0.00	21.00
AE_POTAT	42356	0.96	1.47	1.00	0.00	21.00
AE_BEANS	42356	1.38	2.15	1.00	0.00	21.00
AE_VEGI	42356	6.02	4.75	6.00	0.00	56.00
AE_SODA	42356	1.93	4.24	0.00	0.00	42.00
AE_JUICE	42356	3.05	3.83	2.00	0.00	28.00
AE_FLAV	42356	1.03	2.86	0.00	0.00	42.00
AE_CAKE	42356	2.09	2.68	1.00	0.00	35.00
AE_FROZ	42356	1.20	1.82	1.00	0.00	21.00

SUMMARY OF MATRIX VARIABLES

AE_FRUIT- # eating fruit

AE_SALAD- # eating salad

AE_FRIES- #eating fries

AE_POTAT- # eating potatoes

AE_BEANS- # eating beans

AE_VEGI- # eating vegetables

SUG- teaspoons of added sugar consumed per day

FV-daily servings of fruits and vegetables

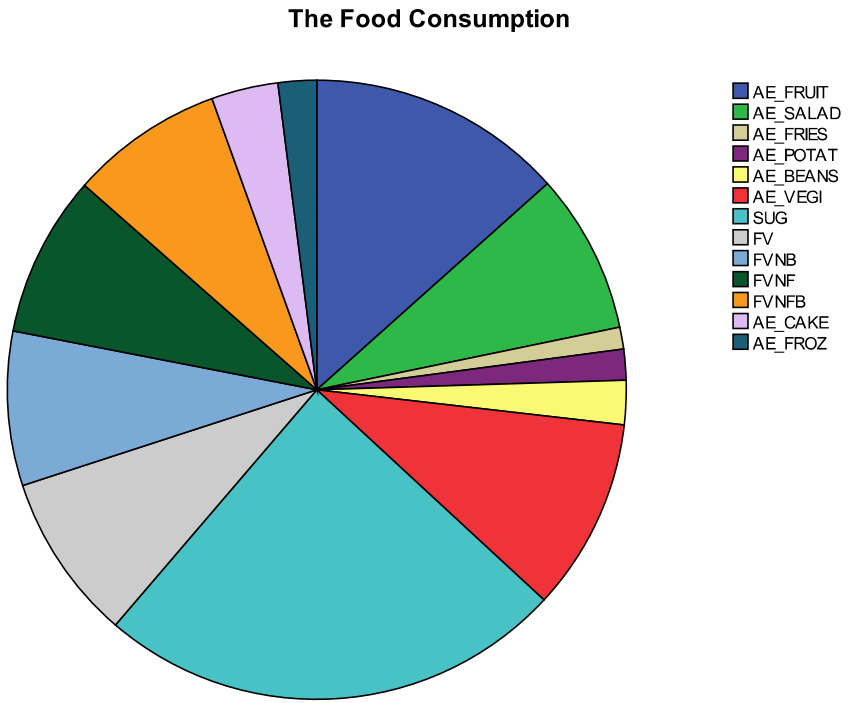
FVNB- daily servings of fruits and vegetables except beans

FVNF- daily servings of fruits and vegetables except French fries

FVNFB- daily servings of fruits and vegetables except French fries and beans

AE_CAKE- number of times # eating cake/pie/cookies per week

AE_FROZ- number of times # eating ice cream/frozen desserts per week



ORDINAL VARIABLES

HEALTH CONDITON

EXCELLENT	21.69%
VERY GOOD	31.94%
GOOD	27.89%
FAIR	13.73%
POOR	4.75%

OVERWEIGHT CONDITION

YES (1)	54.95
NO (0)	45.05

WALKING FOR FUN/EXERCISE

YES (1)	76.94%
NO (0)	23.06%

CANONICAL CORRELATION

- Canonical correlation predicts dependent measures: height and weight from the independent measures: eating fruit, salad, fries, potatoes, beans, vegetables, sugar, drinking soda, drinking juice, drinking flavored drinks, eating cake/pie/cookies and eating ice cream/frozen desserts on the weekly basis.

CANONICAL CORRELATION FUNCTIONS

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.615458	0.615271	0.003018	0.378789
2	0.142295	0.141164	0.004761	0.020248

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
0.6098	0.5891	0.9672	0.9672	0.60863280	745.69	32	84676	<.0001
0.0207		0.0328	1.0000	0.97975224	58.33	15	42339	<.0001

CANONCIAL CORRELATION CONCLUSION

- In terms of practical significance 1st Canonical Function should be interpreted. There is 70.14% of variance being explained by dependent variate.The canonical loadings of each variable to the canonical function show that the highest contribution to the measure of height and weight has the amount of teaspoons of added sugar consumed per day as its canonical loading equals to 0.2437. The next highest one with loading of 0.1561 is the consumption of daily servings of fruits and vegetables. The last two highest ones are: number of times # eating fries per week and daily servings of fruits and vegetables except French fries with loadings of 0.1291 and 0.1271 respectively. The second canonical function regardless of it's statistically significance level does not possess the high loadings on none of the original variables.
- As result of the **Canonical Correlation**, it was possible to obtain the highest correlation of 0.6155 between the height and weight combined together and the consumption on the weekly basis of: fruits, salad, fries, potatoes, beans, vegetables, sugar, drinking soda, drinking juice, drinking flavored drinks, eating cake/pie/cookies and eating ice cream/frozen desserts.

FACTOR ANALYSIS WITH METHOD PRINCIPAL COMPONENTS

- Due to presence of multicollinearity among the food consumption variables, the **Factor Analysis with method Principal Components** as the data reduction method will be applied. The desire is to get rid of the multicollinearity and identify interrelated sets of variables that are uncorrelated. The rotated Factor Analysis can not be performed as the communality is greater than 1. In Factor Analysis not necessary that 100% of variance be accounted for by the extracted factors. The newly created variables (factors) explain shared variation in the original variables.

DERIVED FACTORS AND THEIR LOADINGS

Eigenvalues of the Correlation Matrix: Total = 16 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	5.41703982	2.97048733	0.3386	0.3386
2	2.44655248	1.27737545	0.1529	0.4915
3	1.16917703	0.12264982	0.0731	0.5645
4	1.04652721	0.05653853	0.0654	0.6300

Factor Pattern

	Factor1	Factor2	Factor3	Factor4
AE_FRUIT	0.75652	-0.17725	0.01051	-0.17997
AE_SALAD	0.50597	-0.24190	0.07165	-0.10726
AE_FRIES	-0.00524	0.47833	-0.10930	0.43682
AE_POTAT	0.23829	0.25875	0.23327	0.69619
AE_BEANS	0.17642	0.19898	-0.35054	0.30047
AE_VEGI	0.57257	-0.21553	0.11366	-0.15258
FV	0.97858	0.10254	-0.07711	0.08267
FVNB	0.99173	-0.01614	-0.00029	-0.02180
FVNF	0.99171	0.02367	-0.05535	0.02197
FVNFB	0.99203	-0.01599	-0.00019	-0.02266
SUG	0.03203	0.93657	0.00791	-0.23832
AE_SODA	-0.04763	0.73275	-0.20898	-0.18210
AE_JUICE	0.49424	0.22763	-0.13602	0.06591
AE_FLAV	0.06415	0.52596	-0.25642	-0.31109
AE_CAKE	0.07013	0.32183	0.65386	-0.01073
AE_FROZ	0.08701	0.34223	0.63027	-0.11867

CONCLUSION OF FACTOR ANALYSIS

- The first factor contains the high loadings of the consumption of fruits and vegetables, so it's name is "Consumption of Fruits and Vegetables". Factor 2 contains high loadings of the consumption of soda, sugar and fruit-flavored drinks and can be called "Consumption of Sugar in Drinks". Factor 3 contains high loadings on variables: the consumption of cake, cookies, ice-cream and frozen desserts and can be called "Consumption of Cakes and Deserts". Factor 4 contains high loadings on variables: the consumption of fries, potatoes and beans, so it's called "Consumption of Potatoes and Beans".
- As a result the data reduction method "Factor Analysis" allowed reducing the number of variables from 16 with present multicollinearity to the 4 uncorrelated new variables that explain 63% of the variance in the original variables.

DISCRIMINANT ANALYSIS

- Which variables discriminate between two groups (1-Overweight-Obese) and (0-Not Overweight) out of: daily servings of fruits and vegetables, daily servings of fruits and vegetables except beans, daily servings of fruits and vegetables except French fires, daily servings of fruits and vegetables except French fries and beans, teaspoons of added sugar consumed per day, height and weight? The Stepwise Discriminant Analysis is applied.
- The key assumptions for deriving the Discriminant Function are multivariate normality of the independent (continuous) variables and equality of covariance matrices for the groups defined by the dependent variable. The multivariate normality is not accessible for the independent variables. The variables are not normally distributed; therefore the log transformations for each independent variable have been applied.
- To perform the classification based on the Discriminant Analysis the sample of 42356 was divided into the model building sample (analysis) and validation sample. The samples were selected based on the Random Sampling, in other words selecting randomly 60 % of the records (25340) for the model building using Stepwise Discriminant Analysis and remaining 40% was used for the validation of Disciminant Classification.

STEPWISE DISCRIMINANT ANALYSIS AND CLASSIFICATION RESULTS

Step	Entered	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	lt_WGHTP_P	0.4688	22365.4	<.0001	0.53115660	<.0001	0.46884340	<.0001
2	lt_HGHTI_P	0.2414	8060.66	<.0001	0.40295979	<.0001	0.59704021	<.0001
3	lt_FVNB	0.0003	6.99	0.0082	0.40284871	<.0001	0.59715129	<.0001
4	lt_FVNF	0.0005	13.33	0.0003	0.40263679	<.0001	0.59736321	<.0001
5	lt_FV	0.0004	10.99	0.0009	0.40246225	<.0001	0.59753775	<.0001
6	lt_SUG	0.0003	6.60	0.0102	0.40235748	<.0001	0.59764252	<.0001
7	lt_FVNFB	0.0001	3.74	0.0532	0.40229810	<.0001	0.59770190	<.0001

Number of Observations and Percent Classified into RECOVRWT

From RECOVRWT	0	1	Total
0	7200 94.08	453 5.92	7653 100.00
1	681 7.27	8682 92.73	9363 100.00
Total	7881 46.32	9135 53.68	17016 100.00
Priors	0.44975	0.55025	

CONCLUSION OF DISCRIMINANT ANALYSIS

- The classification of respondents into two groups: overweight and not overweight based on the discriminating variables: weight, height, amount of consumed sugar on the weekly basis, daily servings of fruits and vegetables except beans, daily servings of fruits and vegetables except French fries, daily servings of fruits and vegetables and daily servings of fruits and vegetables except French fries and beans gives an error rate of 6.66%.
- The hit ratio (percent of correctly classified) is the predictive accuracy of the discriminant function. It is measured by (7200 + 8682) =15882/17016 * 100%=93.34%. It can be said that the discriminating variables have overall high power to classify respondents into overweight or not overweight.

LOGISTIC REGRESSION

- The Discriminant Analysis results show the rate of 93.34% correctly classified observations, however as specified despite the log transformation of the discriminating variables, the assumption of multivariate normality was still not met.
- The decision was to use the Logistic Regression with the original variables (not log transformed) plus the age of respondent. The response variable is overweight or not overweight.
- The purpose is to compare the classification results of Discriminant Analysis with those from Logistic Regression

LOGISTIC REGRESSION SUMMARY

- **The Stepwise Logistic Regression was applied to the whole sample without dividing data into analysis and validation sample. As the result of the Stepwise Logistic Regression only height and weight turned out to be the predictors of the response variable**

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	335.0	12.4988	718.4196	<.0001
HGHTI_P	1	-10.2237	0.3730	751.2319	<.0001
WGHTP_P	1	2.1902	0.0783	781.5417	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
HGHTI_P	<0.001	<0.001	<0.001
WGHTP_P	8.937	7.665	10.420

out		Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	Correct 0	18968	44.78	18968	44.78
1	Correct 1	23146	54.65	42114	99.43
2	0 classified 1	114	0.27	42228	99.70
3	1 classified 0	128	0.30	42356	100.00

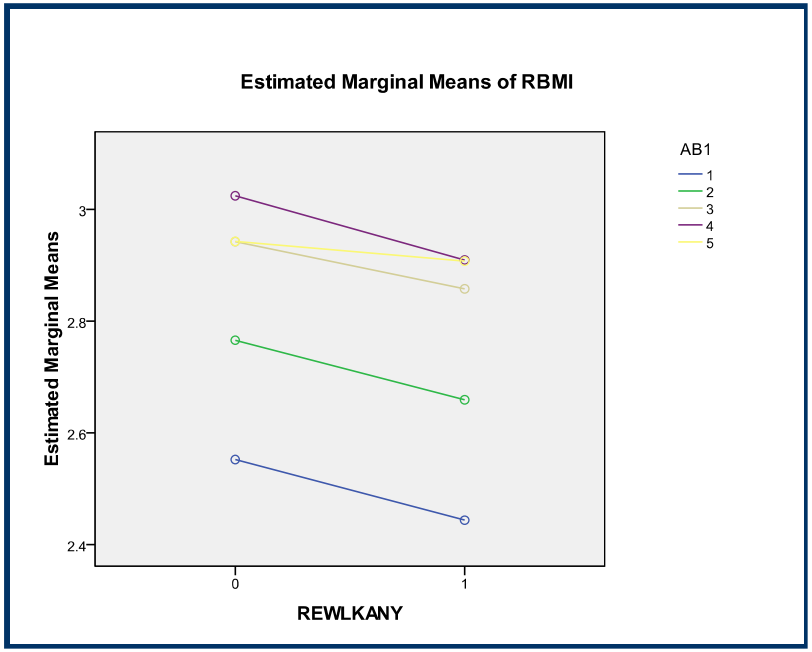
LOGISTIC REGRESSION CONCLUSION

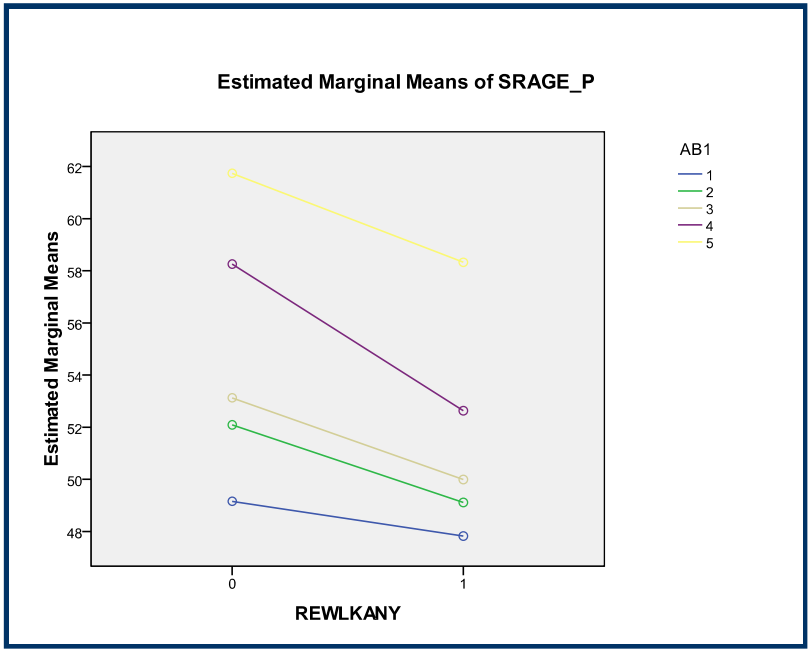
- The odds ratio results show weight is 8.937 times more likely than height to predict whether the person is overweight or not overweight.
- The results presented in the table show that 99.43% of observations are being classified correctly based on the height and weight of respondents.
- The Logistic Regression Stepwise method shows that other variables: the consumption of fruits and vegetables as well as age do not contribute significantly to the prediction of whether respondent is obese (overweight) or not.
- The classification results of Logistic Regression are higher from those of the Discriminant Analysis.

MANOVA MULTIVARIATE METHOD

- Does health condition and exercise/walking combined together have significant effect on the prediction of group differences in body mass index (BMI) and respondent's age?
- The BMI is the ratio of weight and height. BMI is an indication whether the respondent is obese with BMI greater than 30 or healthiest with BMI between 18.5 and 25.
- Among several of assumptions of Manova one of them is normality. Therefore, BMI- body mass index has been transformed by using the log transformation. Another assumption of MANOVA is the homogeneity of variance/covariance matrices. The hypothesis testing applied is:
- Ho: The observed covariance matrices of the dependent variables are equal across the groups
- Ha: The observed covariance matrices of the dependent variables are not equal
- Based on the Box's Test of Equality of Covariance Matrices, the Ho is rejected

OVERALL DIFFERENCES IN AGE AND BMI





AB1: Health Condition: 1-Excellent, 2-Very Good, 3-Good, 4-Fair, 5-Poor

REWLKANY: 0-No walking/exercising 1-walking/exercising

CONCLUSION OF MANOVA

- The highest BMI is for the fair health condition and no exercise or walking. The BMI-body mass index goes down with walking or exercising. The lowest BMI is for excellent health condition and exercising or walking. The BMI goes up when the health condition gets worst.
- The not well health condition and no exercise or walking is for the older age. The youngest age bracket is for excellent health condition and walking or doing exercise for fun. With the increasing age the health condition gets worst. The age bracket between 52 and 50 has Very good health condition. Lower age bracket appears to perform more exercise and walking for fun.
- There is statistically significant difference between the groups with excessive walking and excellent condition and no walking with excellent condition. Also, highly statistical difference exists between the groups with excessive walking with poor health condition and no walking with poor health condition as in that instance the age goes up.

CONCLUSION AND RECOMENDATIONS

Based on the performed statistical analysis, the habitual consumption of food by the individuals has an enormous impact on their body measures: weight and height. Eating salads, beans and fruit and vegetables affects the weight and height in the positive way. On the contrary, the consumption of sugar, fries and potatoes increases the body measures.

The consumption of fruit and vegetables plus height and weight has the power of predicting whether the individual is overweight or not.

Furthermore, the combined effect of health condition and walking/exercising has the statistically significant effect on the differences among the groups in body mass index as well as age. It has been shown that the increasing gradually age has the tendency to decrease the amount of walking and exercising combined with the change of health condition from excellent or good to fair or poor.

Based on the made conclusions, the suggestion for the future research would be to consider the 4 factors obtained from Factor Analysis and apply Logistic Regression or even repeat the Canonical Correlation Analysis to see their affect on the body measures.

REFERENCES

- “Categorical Data Analysis Using Logistic Regression” –Instructor based training Book Code 60444 Course Notes
- “Multivariate Data Analysis” by Joseph F. Hair, Wiliam C. Black, Barry J. Babin, Rolph E. Anderson and Ronald L. Tatham Sixth Edition
- “Simultaneous Surveys of Food Consumption in Various Camps of the United States Army “ by HENEY CLARK SCHOR AND HARMON L.SWAIN
- “Women’s Health And Wellness” by Carol Angstadt, Edited by Prevention, 2002