PROJECT OVERVIEW

Each year, the Center for Disease Control (CDC) conducts a survey known as the Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS is a health-related telephone survey for US residents that collects information about their health conditions and habits. For this project, I chose to analyze the effect between a person's diet / consumption habits and their health. I wanted to see if a person's health could predict their diet, or vice versa.

DATA ANALYSIS PROCESS OVERVIEW

Analyzing the data first required it to be filtered and cleaned. This would ensure that only essential variables were in the dataset, and that these variables were in a format that was easy to understand and visualize. The variables examined were:

- **Identifiers:** The respondent's sex (male or female)
- **Overall Health:** The respondent's overall health, physical health, and mental health
- **Health Issues:** Whether or not the respondent had high blood pressure, high cholesterol, a heart attack, coronary heart disease, stroke, asthma, skin cancer, another type of cancer, lung diseases, arthritis, depressive disorder, kidney disease, or diabetes
- **Diet Variables:** The alcohol drinking habits of the respondent (average number of drinks, the number of days drinking, etc.), the respondent's consumption of fruit, pure fruit juice, vegetables (green and non-green), potatoes (fried and non-fried), soda, sugar-sweetened drinks, the respondent's reduction of sodium, salt, and alcohol for their diet, and the ability to afford balanced meals

To analyze the data, I first grouped the diet variables into numerical and categorical predictors. Then, I made two SAS macros:

1. **compare_health_var_num**: This macro takes the input of a list of categorical variables and a list of numerical predictors, and creates three separate outputs – a PROC MEANS table, a PROC SGPLOT vertical box (vbox) plot, and a PROC NPAR1WAY

for comparing the differences between two categories (either "yes" and "no", or the highest and lowest category for a dataset).

2. **compare_health_var_cat**: This macro takes the input of two lists of categorical variables, and creates a PROC FREQ output comparing the proportions of the two datasets.

I then ran the macros, using various inputs, to try and find trends in the data.

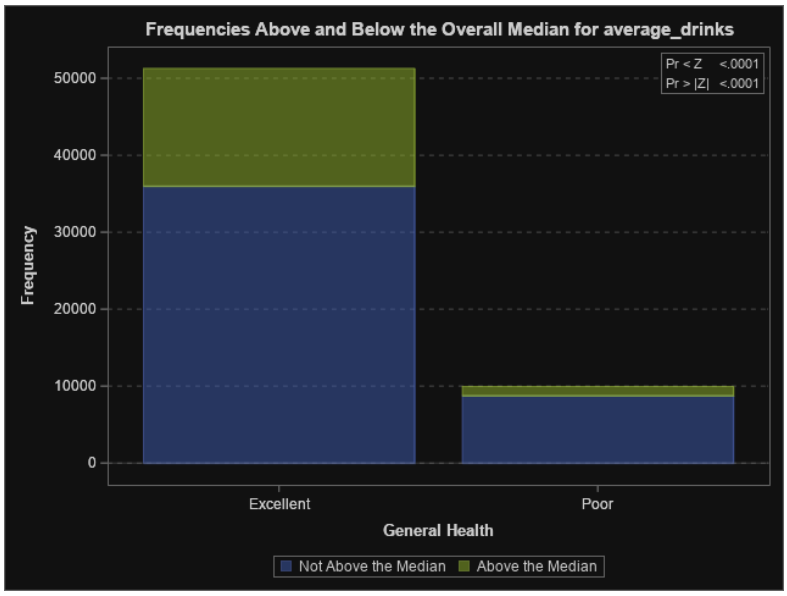
After this, I then decided to find the predictors from the dataset that could best predict a person's health. I created two models: one for a person's physical health, and one for a person's mental health. For both models, I started by finding which variables were most significant on the model using PROC GLM. I then used stepwise model selection (thru PROC GLMSELECT) to find the most significant variables for the final model.

DATA ANALYSIS INSIGHTS

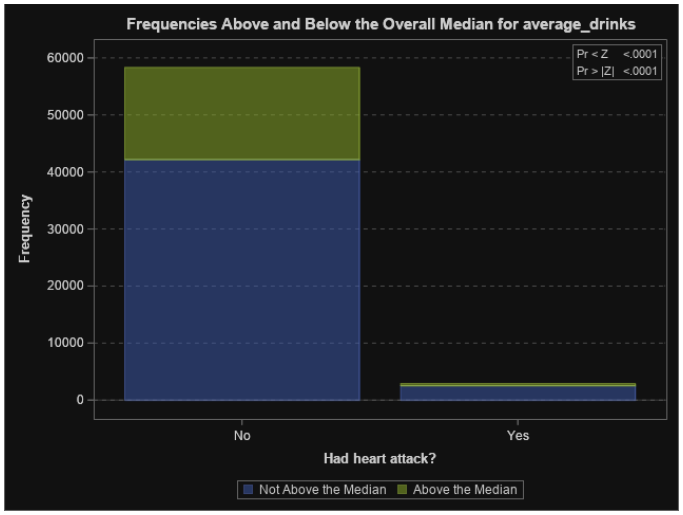
From running these plots, we can see some insights:

1. **People who are in good health drink more alcohol and more frequently than those in poor health.** In terms of general, mental, and physical health, it can be seen that healthy people consume more alcohol than non-healthy people. Additionally, those who have had <u>any</u> major health issue (heart attack, stroke, etc.) consumed less alcohol on average. An example of this is demonstrated by following figures:

Analysis Variable : average_drinks Average drinks per day							
General Health	N Obs	N	Minimum	Maximum	Mean	Median	Std Dev
Excellent	55321	55321	0	16.0000000	1.1532691	1.0000000	1.4752197
Very Good	111745	111745	0	17.0000000	1.1517025	1.0000000	1.4642930
Good	102987	102987	0	16.0000000	0.9963199	0	1.5635590
Fair	42860	42860	0	16.0000000	0.7636024	0	1.5087480
Poor	16124	16124	0	16.0000000	0.5233193	0	1.2935038



Analysis Variable : average_drinks Average drinks per day							
Had heart attack?	N Obs	N	Minimum	Maximum	Mean	Median	Std Dev
No	310829	310829	0	17.0000000	1.0433904	1.0000000	1.5162626
Yes	18208	18208	0	16.0000000	0.6565795	0	1.2595675



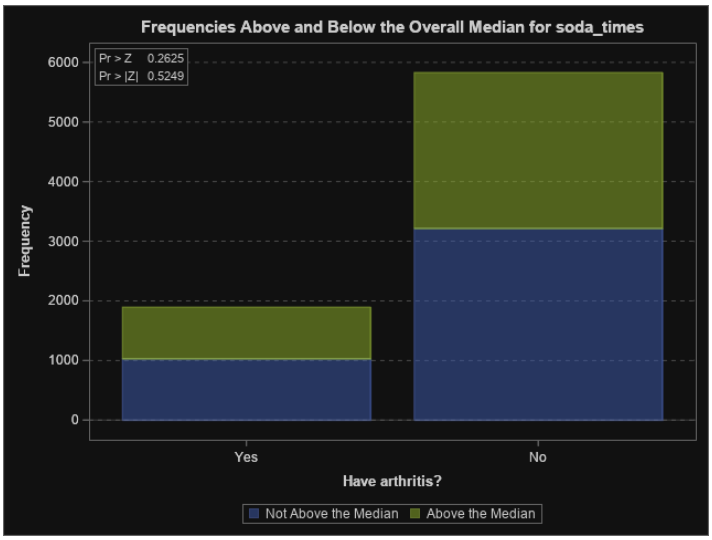
2. **People who are in good health eat more fruits and vegetables than those in poor health.** In terms of general, mental, and physical health, healthier people reported eating more fruits and vegetables (green and non-green) than non-healthy people. However, potatoes (non-fried and fried) and juices seemed to have no effect. An example of this is demonstrated by following figures:

Analysis Variable : veggie_times Average green vegetable consumption frequency per day							
Number of days mental health was not good in past 30 days (grouped)	N Obs	N	Minimum	Maximum	Mean	Median	Std Dev
1–5 days	277977	274043	0	15.0000000	0.6058159	0.4285714	0.6261790
6–10 days	15626	15420	0	10.0000000	0.5378519	0.4285714	0.5884580
11–15 days	10390	10239	0	15.0000000	0.5267051	0.4285714	0.6379726
16–20 days	5146	5080	0	10.0000000	0.5072732	0.3333333	0.6152455
21–25 days	2108	2076	0	10.0000000	0.4797917	0.2857143	0.5876477
26–30 days	17790	17518	0	15.0000000	0.4936183	0.3333333	0.6192098

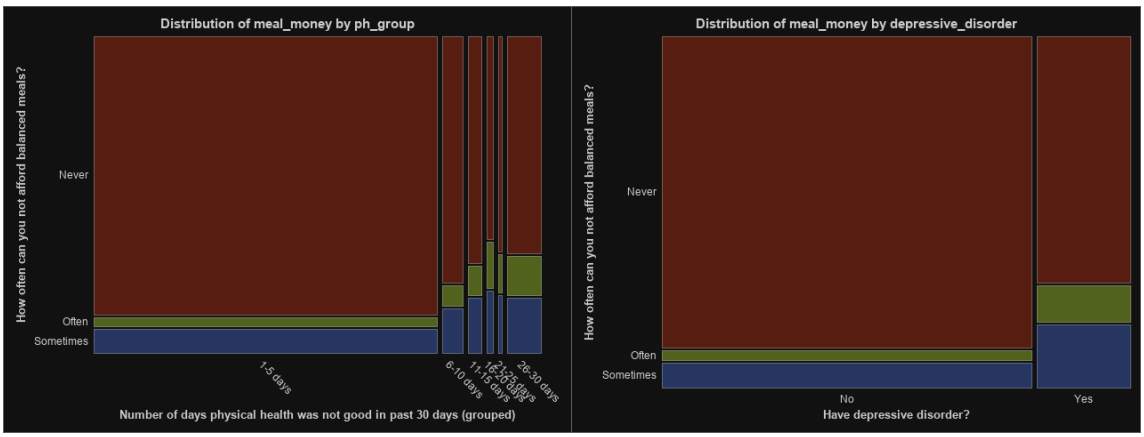
Analysis Variable : fry_times Average fried potato consumption frequency per day							
Number of days mental health was not good in past 30 days (grouped)	N Obs	N	Minimum	Maximum	Mean	Median	Std Dev
1–5 days	277977	273632	0	15.0000000	0.1905466	0.1428571	0.2919431
6–10 days	15626	15394	0	10.0000000	0.2219993	0.1428571	0.3243708
11–15 days	10390	10221	0	10.5714286	0.2266473	0.1428571	0.3528945
16–20 days	5146	5072	0	9.0000000	0.2269172	0.1428571	0.3287269
21–25 days	2108	2073	0	10.0000000	0.2257804	0.1428571	0.3741829
26–30 days	17790	17483	0	10.0000000	0.2157252	0.1333333	0.3448184

3. **People who are in good health drink less soda and sugar-sweetened drinks than those in poor health.** In terms of general, mental, and physical health, people who reported being in good health drank less soda and sugar-sweetened drinks than those in poor health. However, there seemed to be no significant difference between a person's soda consumption and any specific health issue. An example of this is demonstrated by following figures:

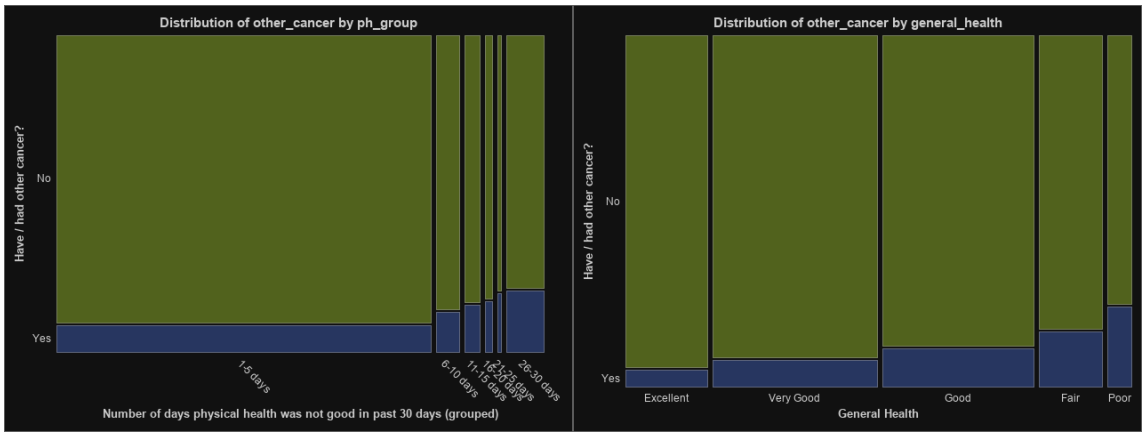
Analysis Variable : soda_times Average soda consumption frequency per day							
General Health	N Obs	N	Minimum	Maximum	Mean	Median	Std Dev
Excellent	55321	6870	0	12.0000000	0.2091093	0	0.5815951
Very Good	111745	13760	0	12.0000000	0.2532485	0.0333333	0.6488167
Good	102987	13140	0	15.0000000	0.3340030	0.0333333	0.7802034
Fair	42860	5335	0	15.0000000	0.4317954	0.0666667	0.9641157
Poor	16124	2131	0	15.0000000	0.4734620	0.0666667	1.1100317

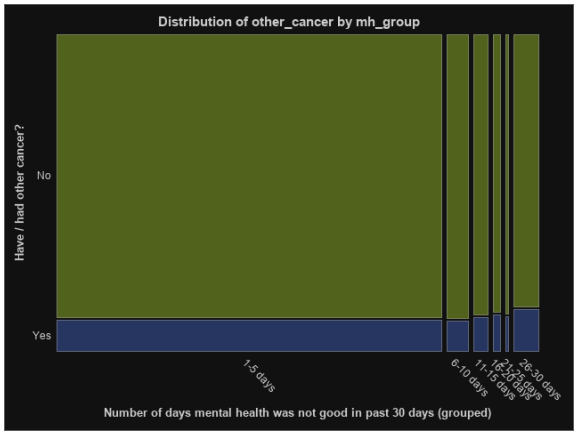


4. **People who could not afford balanced meals were more likely to report they had worse health.** In terms of general, mental, and physical health, those who could not afford balanced meals were more likely to report that they had worse health. This also applied to the various health issues. An example of this is demonstrated by following figures:



5. **People who had a health issue were more likely to report poor general and physical health.** Respondents who had a previous health issue or condition were more likely to report having lower general health than those who hasn't. However, the mental health was not generally affected by this, with the exception of depressive disorder. An example of this is demonstrated by following figures:

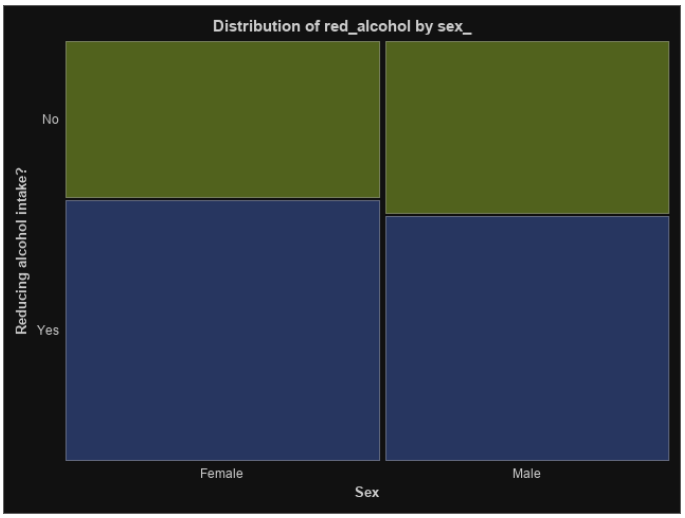




6. **Females appear to be slightly healthier than males.** While respondents' general health was not affected by gender, females generally had healthier eating habits. This includes consuming more fruits and vegetables on average, drinking less alcohol and soda on average, and reducing unhealthy consumption more, on average. An example of this is demonstrated by following figures:

Analysis Variable : fruit_times Average fruit consumption frequency per day							
Sex	N Obs	N	Minimum	Maximum	Mean	Median	Std Dev
Female	189562	187730	0	15.0000000	1.2357330	1.0000000	1.0301480
Male	139475	137991	0	15.0000000	1.0269952	1.0000000	0.9543042

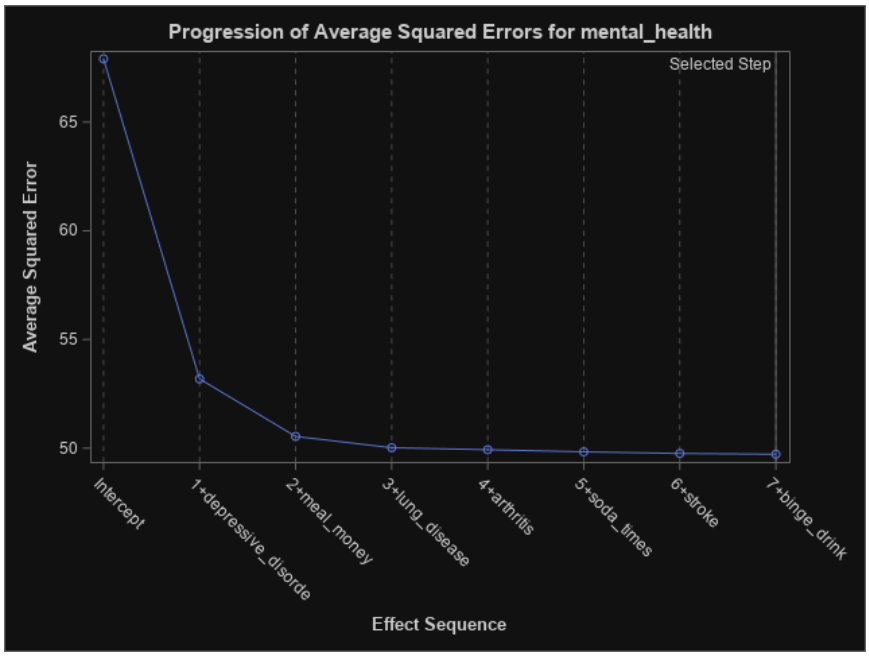
Analysis Variable : average_drinks Average drinks per day							
Sex	N Obs	N	Minimum	Maximum	Mean	Median	Std Dev
Female	189562	189562	0	16.0000000	0.7932128	0	1.2083572
Male	139475	139475	0	17.0000000	1.3329127	1.0000000	1.7879805



MODELING INSIGHTS

For the **mental health model**, the best predictors were the presence of depressive disorder, the ability to afford balanced meals, and the presence of a lung disease. This can be seen from the summary figures of the stepwise selection:

Stepwise Selection Summary for Mental Health										
Step	Effect Entered	Model R-Square	Adjusted R-Square	AIC	BIC	SBC	PRESS	ASE	F Value	Pr > F
0	Intercept	0.0000	0.0000	41222.3494	33322.6728	33328.3239	536551.469	67.9093	0.00	1.0000
1	depressive_disorder	0.2168	0.2167	39294.1277	31394.8630	31407.0767	420453.435	53.1868	2185.95	<.0001
2	meal_money	0.2558	0.2555	38894.7836	30995.6624	31021.6816	399943.971	50.5391	206.80	<.0001
3	lung_disease	0.2634	0.2631	38815.1235	30916.0748	30948.9959	396003.084	50.0193	82.03	<.0001
4	arthritis	0.2648	0.2643	38802.7145	30903.6780	30943.5614	395385.313	49.9282	14.41	0.0001
5	soda_times	0.2662	0.2657	38789.3545	30890.3390	30937.1759	394757.463	49.8312	15.36	<.0001
6	stroke	0.2673	0.2667	38779.5746	30880.5809	30934.3706*	394363.038	49.7569	11.78	0.0006
7	binge_drink	0.2679	0.2671*	38775.6526*	30876.6732*	30937.4231	394199.980*	49.7196	5.92	0.0150
* Optimal Value of Criterion										

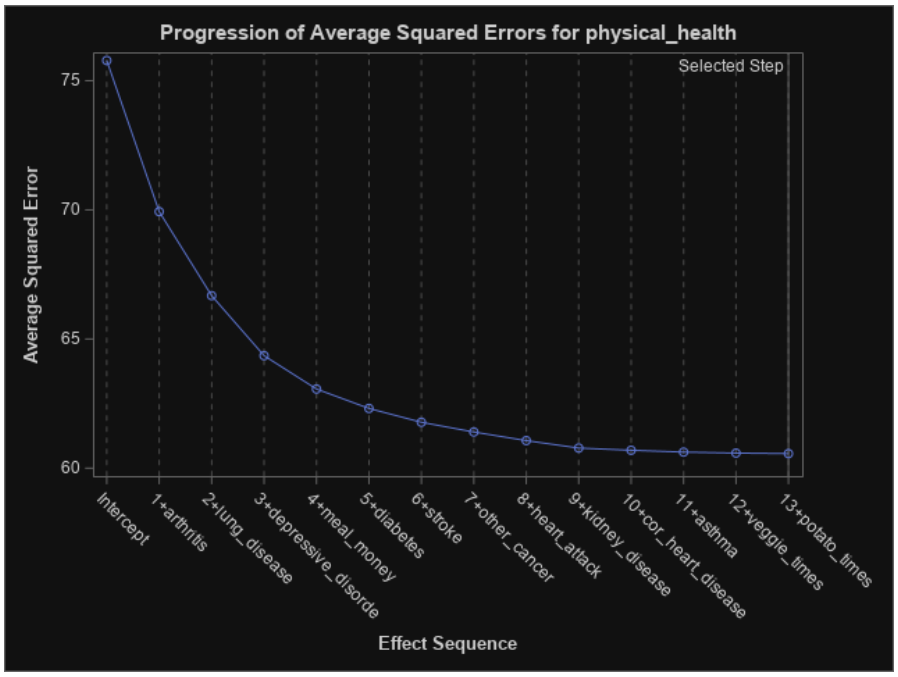


For the **physical health model**, the best predictors selected were the presence of arthritis, the presence of a lung disease, the presence of depressive disorder, the ability to afford balanced meals, the presence of diabetes, the presence of stroke, the current / previous presence of cancer, if the participant ever had a heart attack, and the presence of kidney disease. This can be seen from the summary figures of the stepwise selection:

Stepwise Selection Summary for Physical Health

Step	Effect Entered	Model R-Square	Adjusted R-Square	AIC	BIC	CP	SBC	PRESS	ASE	F Value	Pr > F
0	Intercept	0.0000	0.0000	357354.107	290281.626	16837.3437	290289.221	5082988.84	75.7818	0.00	1.0000
1	arthritis	0.0772	0.0772	351967.123	284894.552	10361.4251	284911.350	4690757.40	69.9312	5611.23	<.0001
2	lung_disease	0.1201	0.1201	348774.030	281701.464	6761.4131	281727.371	4472820.71	66.6780	3272.27	<.0001
3	depressive_disorder	0.1508	0.1508	346396.263	279323.787	4189.8845	279358.717	4317094.99	64.3537	2422.34	<.0001
4	meal_money	0.1678	0.1678	345040.780	277968.304	2764.1768	278021.461	4230878.16	63.0624	686.62	<.0001
5	diabetes	0.1777	0.1777	344239.729	277167.337	1935.1586	277229.523	4180724.00	62.3119	807.79	<.0001
6	stroke	0.1848	0.1847	343662.797	276590.485	1344.2068	276661.705	4145036.23	61.7764	581.37	<.0001
7	other_cancer	0.1898	0.1897	343253.674	276181.433	928.2255	276261.696	4119906.56	61.3989	412.33	<.0001
8	heart_attack	0.1942	0.1941	342891.276	275819.115	561.8830	275908.411	4097809.05	61.0662	365.34	<.0001
9	kidney_disease	0.1980	0.1979	342576.246	275504.173	245.0460	275602.495	4078726.69	60.7782	317.73	<.0001
10	cor_heart_disease	0.1991	0.1990	342482.960	275410.915	151.5116	275518.323	4073154.90	60.6920	95.34	<.0001
11	asthma	0.2001	0.1999	342407.014	275334.995	75.4625	275451.490	4068583.89	60.6215	77.98	<.0001
12	veggie_times	0.2006	0.2004	342366.472	275294.469	34.9028	275420.062	4066136.06	60.5830	42.55	<.0001
13	potato_times	0.2008	0.2007*	342345.723*	275273.730*	14.1550*	275408.426*	4064890.85*	60.5625	22.75	<.0001

* Optimal Value of Criterion



For both models, we can see in the figures (through factors like AIC, BIC, and the R-Squared Value) that the predictors listed above had the most drastic effect on the models; the other predictors had a less significant effect.

Comparing the mental health and physical health models, there are three interesting observations:

1. The effects of the predictors in the mental health model are more drastic than the effects of the predictors in the physical health model. That is, the predictor variables had much more of a gradual effect on the physical health model.
2. The mental and physical health models both had the same top four variables: the presence of arthritis, the presence of a lung disease, the presence of depressive disorder, and the ability to afford balanced meals.
3. Most of the top predictors were health variables, meaning that a person's health is more affected by their conditions than their diet.

Just out of curiosity, I ran PROC GLMSELECT using only dietary variables, and found they had a much less significant impact on both the mental and physical health models than the health-related predictors. This can be seen below with the significantly-reduced R-Squared Values:

Stepwise Selection Summary for Physical Health											
Step	Effect Entered	Model R-Square	Adjusted R-Square	AIC	BIC	CP	SBC	PRESS	ASE	F Value	Pr > F
0	Intercept	0.0000	0.0000	221815.841	180625.792	1012.1628	180632.467	3305469.06	80.2454	0.00	1.0000
1	days_drank	0.0108	0.0107	221371.795	180181.741	559.6432	180197.047	3269992.63	79.3811	448.45	<.0001
2	veggie_times	0.0140	0.0140	221237.613	180047.553	423.8527	180071.491	3259360.48	79.1191	136.40	<.0001
3	max_drinks	0.0159	0.0159	221160.481	179970.416	345.9944	180002.985	3253247.66	78.9673	79.20	<.0001
4	binge_drink	0.0179	0.0178	221079.241	179889.180	264.1563	179930.371	3246840.71	78.8078	83.31	<.0001
5	potato_times	0.0196	0.0195	221008.403	179818.351	192.9336	179868.159	3241321.73	78.6686	72.89	<.0001
6	fruit_times	0.0215	0.0214	220931.435	179741.401	115.6947	179799.817	3235268.64	78.5179	79.03	<.0001
7	swtdrnk_times	0.0228	0.0226	220879.891	179689.872	64.0528	179756.898	3231294.15	78.4159	53.57	<.0001
8	soda_times	0.0235	0.0233	220852.334	179662.326	36.4721	179737.967	3229196.90	78.3597	29.56	<.0001
9	ngveggie_times	0.0238	0.0236	220841.367	179651.364	25.5011	179735.626	3228329.26	78.3350	12.97	0.0003
10	average_drinks	0.0241	0.0238	220831.856	179641.859	15.9903	179734.741*	3227556.09	78.3131	11.51	0.0007
11	juice_times	0.0242	0.0240*	220827.311*	179637.318*	11.4468*	179738.822	3227216.83*	78.3007	6.54	0.0105
* Optimal Value of Criterion											