Regression Analysis Student Project
California's Health Analysis
Summer Session July 15th – Sept 6th

Ilina Sandler

Objective:
My objective in this study is to examine the relationship between adult obesity and a selection of lifestyle and behavioral choices, yielding a model with the best fit. The dependent variable is adult obesity (depicted as a percentage of the population) and there are nine independent variables: adult smoking, physical inactivity, percentage of excessive drinking, percent uninsured, number of primary care physicians, dentists, high school graduation, percent unemployed, number of fast food restaurants. I will perform Ordinary Least Squares Regression, using a 95% confidence interval.

Data:
The data can be seen below, and was obtained from http://www.countyhealthrankings.org/app/california/2013/compare-counties/. I chose thirteen counties to analyze:

	Adult obesity	Adult smoking	Physical inactivity	Excessive drinking	Uninsured	Primary care physicians	Dentists	High school graduation	Unemployment	Fast food restaurants
Contra Costa (CN)	**24%**	13%	18%	19%	14%	1,121	1,364	83%	10.40%	47%
Imperial (IM)	**25%**	11%	23%	15%	23%	4,170	3,318	85%	29.70%	50%
Tuolumne (TO)	**23%**	21%	19%	26%	16%	1,062	1,218	89%	13.00%	31%
Yuba (YU)	**31%**	14%	25%	16%	19%	3,618	4,902	81%	18.20%	57%
Napa (NA)	**22%**	9%	16%	23%	19%	1,190	1,492	86%	9.00%	37%
San Francisco (SF)	**17%**	11%	17%	21%	15%	688	805	82%	8.60%	30%
Los Angeles (LO)	**22%**	13%	19%	16%	26%	1,415	1,402	75%	12.30%	50%
San Luis Obispo (SP)	**22%**	10%	15%	20%	18%	1,280	1,465	92%	9.30%	42%
Santa Cruz (SC)	**20%**	10%	12%	18%	18%	1,047	1,547	86%	12.10%	42%
Humboldt (HU)	**26%**	20%	19%	21%	20%	1,334	1,644	86%	11.30%	40%
Placer (PL)	**20%**	10%	14%	17%	12%	929	1,096	91%	10.80%	48%
Ventura (VE)	**23%**	12%	17%	18%	18%	1,458	1,386	86%	10.10%	49%
Santa Barbara (SR)	**20%**	11%	16%	18%	21%	1,253	1,460	86%	8.80%	45%

Equation:

For the analysis, I will use the following equation,

$$Y = A + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5 + B_6X_6 + B_7X_7 + B_8X_8 + B_9X_9$$

The dependent variable (Y) is the calories per each sandwich. A is the intercept, and the independent variables (the X's) are listed below:

X_1= Adult smoking (%)
X_2= Physical inactivity (%)
X_3= Excessive drinking (%)
X_4= Uninsured (%)
X_5= Primary care physicians (#)
X_6= High school graduation (%)
X_7= Unemployment (%)
X_8= Fast food restaurants (%)
X_9= Dentists (#)

Models:

In order to estimate the parameters of the following models, I used the regression function in Excel's Data Analysis Add-In. First, the full model is evaluated, including all nine explanatory variables. Based on the results, variables will be eliminated one by one- starting with the variable that has the highest p-value.

Model 1- The Full Model

Utilizing all nine explanatory variables in the regression yields the following results;

ALL 9 VARIABLES								
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.97566							
R Square	0.95191							
Adjusted R Square	0.80764							
Standard Error	0.01513							
Observations	13							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	9	0.01359	0.00151	6.59820	0.07393			
Residual	3	0.00069	0.00023					
Total	12	0.01428						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.0886	0.3839	0.2309	0.8323	-1.1330	1.3103	-1.1330	1.3103
X1= Adult smoking (%)	0.4597	0.3574	1.2862	0.2886	-0.6777	1.5972	-0.6777	1.5972
X2= Physical inactivity (%)	-0.4226	0.8073	-0.5235	0.6368	-2.9918	2.1466	-2.9918	2.1466
X3= Excessive drinking (%)	0.7728	0.4555	1.6965	0.1884	-0.6768	2.2224	-0.6768	2.2224
X4= Uninsured (%)	-0.1400	0.2764	-0.5066	0.6473	-1.0197	0.7396	-1.0197	0.7396
X5= Primary care physicians (#)	0.0000	0.0001	0.6754	0.5478	-0.0002	0.0003	-0.0002	0.0003
X6= High school graduation (%)	-0.1764	0.3164	-0.5575	0.6161	-1.1835	0.8307	-1.1835	0.8307
X7= Unemployment (%)	-0.3119	0.4957	-0.6291	0.5739	-1.8895	1.2657	-1.8895	1.2657
X8= Fast food restaurants (%)	0.3344	0.1594	2.0984	0.1268	-0.1727	0.8415	-0.1727	0.8415
X9= Dentists (#)	0.0000	0.0000	0.1130	0.9172	-0.0001	0.0001	-0.0001	0.0001

The initial regression on all nine independent variables produces this equation:

$$Y = 0.0886 + 0.4597_1 - 0.4226X_2 + 0.7728X_3 - 0.1400X_4 + 0.0000X_5 - 0.1764X_6 - 0.3119X_7 - 0.3344X_8 + 0.0000X_9$$

The R^2 of this regression is 0.9519, the Adjusted R^2, which adjusts for the degrees of freedom, is 0.8076. Since these values are relatively close to 1, we see that this model is an adequate fit to the data. The F ratio is 6.5982, and the Signifiance F is 0.0739.

In order to see if we can get a closer fit, we need to analyze the P-values. Small P-values indicate that these explanatory variables really influence the response variable. Therefore, we can remove the explanatory variable with the highest P-value and rerun the regression to make our model more precise. In this case, we rerun the regression without X_9=Dentists, which has a P-value of 0.9172.

Model 2- Eight Explanatory Variables

Removing the number of dentists from the regression and utilizing the other eight explanatory variables yields the following results;

8 VARIABLES								
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.9756							
R Square	0.9517							
Adjusted R Square	0.8551							
Standard Error	0.0131							
Observations	13							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	8	0.0136	0.0017	9.8533	0.0212			
Residual	4	0.0007	0.0002					
Total	12	0.0143						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.1211	0.2206	0.5491	0.6122	-0.4913	0.7336	-0.4913	0.7336
X1= Adult smoking (%)	0.4896	0.2089	2.3433	0.0791	-0.0905	1.0696	-0.0905	1.0696
X2= Physical inactivity (%)	-0.4908	0.4649	-1.0558	0.3506	-1.7816	0.8000	-1.7816	0.8000
X3= Excessive drinking (%)	0.7867	0.3806	2.0668	0.1076	-0.2701	1.8434	-0.2701	1.8434
X4= Uninsured (%)	-0.1641	0.1525	-1.0763	0.3424	-0.5876	0.2593	-0.5876	0.2593
X5= Primary care physicians (#)	0.0001	0.0000	2.3305	0.0802	0.0000	0.0001	0.0000	0.0001
X6= High school graduation (%)	-0.2032	0.1820	-1.1161	0.3269	-0.7086	0.3022	-0.7086	0.3022
X7= Unemployment (%)	-0.3593	0.2287	-1.5708	0.1913	-0.9944	0.2758	-0.9944	0.2758
X8= Fast food restaurants (%)	0.3340	0.1383	2.4157	0.0731	-0.0499	0.7179	-0.0499	0.7179

The second regression on eight independent variables produces this equation:
$Y = 0.1211 + 0.4896X_1 - 0.4908X_2 + 0.7867X_3 - .1641X_4 + .0001X_5 - 0.2032X_6 - .3593X_7 + 0.3340X_8$

With this regression, the R^2 is .9517, almost the exact same as the full model. The Adjusted R^2 is .8551, which is a tiny bit larger than that of the full model. The standard error of Model 2 is also less than Model 1's standard error (0.0131 vs. 0.0151). Lastly, when comparing the F ratio, we can also see that Model 2 has an F ratio of 9.8533, which is larger than Model 1. Based on this analysis, Model 2 may be an better fit to the data than Model 1. The next step, would be to remove the explanatory variable with the highest P-value, Physical inactivity (%) (X_2), to see if that will further enhance the model.

Model 3- Seven Explanatory Variables

Removing physical inactivity from Model 2 and rerunning the regression yields the following results;

7 VARIABLES								
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.9686							
R Square	0.9382							
Adjusted R Square	0.8518							
Standard Error	0.0133							
Observations	13							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	7	0.0134	0.0019	10.8527	0.0092			
Residual	5	0.0009	0.0002					
Total	12	0.0143						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.0581	0.1424	-0.4083	0.7000	-0.4243	0.3080	-0.4243	0.3080
X1= Adult smoking (%)	0.3295	0.1454	2.2661	0.0728	-0.0443	0.7033	-0.0443	0.7033
X3= Excessive drinking (%)	0.6988	0.3756	1.8602	0.1219	-0.2668	1.6644	-0.2668	1.6644
X4= Uninsured (%)	-0.0836	0.1336	-0.6260	0.5588	-0.4270	0.2597	-0.4270	0.2597
X5= Primary care physicians (#)	0.0000	0.0000	2.6251	0.0468	0.0000	0.0001	0.0000	0.0001
X6= High school graduation (%)	-0.0517	0.1134	-0.4564	0.6672	-0.3432	0.2397	-0.3432	0.2397
X7= Unemployment (%)	-0.2560	0.2091	-1.2243	0.2754	-0.7936	0.2816	-0.7936	0.2816
X8= Fast food restaurants (%)	0.3436	0.1395	2.4627	0.0570	-0.0151	0.7023	-0.0151	0.7023

The third regression on seven independent variables produces this equation:
$Y = -0.0581 + 0.3295X_1 + 0.6988X_3 - 0.0836X_4 - 0X_5 - 0.0517X_6 - 0.256X_7 + 0.3436X_8$

The R^2 decreased 0.0135 from Model 2 to Model 3, it is 0.9382. Adjusted R^2 also decreased to 0.8518, the F ratio increased to 10.8527, and the Standard Error increased to 0.0133. Hence, we can deduce that Model 3 is not an improvement from Models 1 and 2.
However, to make sure, we should analyze the results we need to continue enhancing the model by pulling out X_5= High school graduation (%) because it's P-value is the highest.

Model 4- Six Explanatory Variables

Removing high school graduation from Model 3 and rerunning the regression yields the following results;

6 VARIABLES								
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.9673							
R Square	0.9357							
Adjusted R Square	0.8714							
Standard Error	0.0124							
Observations	13							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	6	0.0134	0.0022	14.5461	0.0024			
Residual	6	0.0009	0.0002					
Total	12	0.0143						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.0902	0.1155	-0.7810	0.4645	-0.3727	0.1924	-0.3727	0.1924
X1= Adult smoking (%)	0.3500	0.1289	2.7156	0.0348	0.0346	0.6654	0.0346	0.6654
X3= Excessive drinking (%)	0.6367	0.3263	1.9515	0.0988	-0.1616	1.4351	-0.1616	1.4351
X4= Uninsured (%)	-0.0574	0.1124	-0.5110	0.6276	-0.3324	0.2176	-0.3324	0.2176
X5= Primary care physicians (#)	0.0000	0.0000	2.9261	0.0264	0.0000	0.0001	0.0000	0.0001
X7= Unemployment (%)	-0.2792	0.1890	-1.4769	0.1902	-0.7417	0.1834	-0.7417	0.1834
X8= Fast food restaurants (%)	0.3297	0.1269	2.5991	0.0407	0.0193	0.6401	0.0193	0.6401

The fourth regression on six independent variables produces this equation:
$Y = -0.0902 + 0.35X_1 + 0.6367X_3 - 0.0574X_4 + 0X_5 - 0.2792X_7 + 0.3297X_8$

The R^2 decreased .0026 in this model to .9357. This time, the Adjusted R^2 increased to .8714, the F ratio increased to 14.5461, and the Standard Error decreased to 0.0124. Model 4 is still a precise fit. To see if we can get a better fit, we must continue enhancing the model by rerunning the regression again, removing X_4= percent of people uninsured, with the highest P-value of 0.6276.

Model 5- Five Explanatory Variables

Removing Trans Fat from Model 4 and rerunning the regression yields the following results;

5 VARIABLES								
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.9659							
R Square	0.9329							
Adjusted R Square	0.8849							
Standard Error	0.0117							
Observations	13							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	5	0.0133	0.0027	19.4567	0.0006			
Residual	7	0.0010	0.0001					
Total	12	0.0143						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.1136	0.1003	-1.1327	0.2947	-0.3506	0.1235	-0.3506	0.1235
X1= Adult smoking (%)	0.3353	0.1188	2.8220	0.0257	0.0543	0.6162	0.0543	0.6162
X3= Excessive drinking (%)	0.6848	0.2955	2.3172	0.0536	-0.0140	1.3835	-0.0140	1.3835
X5= Primary care physicians (#)	0.0000	0.0000	3.0596	0.0183	0.0000	0.0001	0.0000	0.0001
X7= Unemployment (%)	-0.2639	0.1765	-1.4949	0.1786	-0.6813	0.1535	-0.6813	0.1535
X8= Fast food restaurants (%)	0.3435	0.1172	2.9309	0.0220	0.0664	0.6207	0.0664	0.6207

The fifth regression on five independent variables produces this equation:
$Y = -0.1136 + 0.3353X_1 + 0.6848X_3 + 0X_5 - 0.2639X_7 + 0.3435X_8$

For Model 5, the R^2 decreased to 0.9329, the Adjusted R^2 increased to 0.8849, and the F ratio increased to 19.4567. The Standard Error decreased to 0.0117. This model is a better fit than Model 4, according to these statistics (although R^2 decreased a bit, Adjusted R^2 increased. To determine if there is a model that is a better fit, we need to run another model without X_7=Unemployment since it has the highest P-value of the explanatory variables left.

Model 6- Four Explanatory Variables

Removing Unemployment from Model 5 and rerunning the regression yields the following results;

4 VARIABLES								
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.9547							
R Square	0.9114							
Adjusted R Square	0.8672							
Standard Error	0.0126							
Observations	13							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	4	0.0130	0.0033	20.5850	0.0003			
Residual	8	0.0013	0.0002					
Total	12	0.0143						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.2034	0.0862	-2.3581	0.0461	-0.4023	-0.0045	-0.4023	-0.0045
X1= Adult smoking (%)	0.2760	0.1203	2.2936	0.0510	-0.0015	0.5534	-0.0015	0.5534
X3= Excessive drinking (%)	0.9035	0.2758	3.2755	0.0113	0.2674	1.5396	0.2674	1.5396
X5= Primary care physicians (#)	0.0000	0.0000	3.9047	0.0045	0.0000	0.0000	0.0000	0.0000
X8= Fast food restaurants (%)	0.4464	0.1020	4.3783	0.0024	0.2113	0.6815	0.2113	0.6815

The sixth regression on four independent variables produces this equation:
$Y = -0.2034 + 0.276X_1 + 0.9035X_3 + 0X_5 - 0.4464X_7$

Model 6 shows a decrease in both R^2 and decrease in Adjusted R^2. Although there is an increase in the F ratio there is also an increase in the Standard Error. From this, I can determine that this model is a worse fit than the previous models.

Just to be sure, it is beneficial to run another regression eliminating X_1= Adult smoking (%).

Model 7- Three Explanatory Variables

Removing Adult smoking (%) from Model 6 and rerunning the regression yields the following results;

3 VARIABLES								
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.9237							
R Square	0.8532							
Adjusted R Square	0.8043							
Standard Error	0.0153							
Observations	13							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	0.0122	0.0041	17.4378	0.0004			
Residual	9	0.0021	0.0002					
Total	12	0.0143						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.2659	0.0993	-2.6767	0.0253	-0.4906	-0.0412	-0.4906	-0.0412
X3= Excessive drinking (%)	1.2452	0.2818	4.4187	0.0017	0.6077	1.8827	0.6077	1.8827
X5= Primary care physicians (#)	0.0000	0.0000	3.7714	0.0044	0.0000	0.0000	0.0000	0.0000
X8= Fast food restaurants (%)	0.5115	0.1189	4.3035	0.0020	0.2426	0.7804	0.2426	0.7804

The seventh regression on three independent variables produces this equation:
$Y = -0.2659 + 1.2452X_3 + 0X_5 + 0.5115X_8$

In this model, R^2, Adjusted R^2, the F ratio have all decreased while the Standard Error has increased. From these results, it is quite clear that this is not the optimal model. Based on the decrease in fit of models 6 and 7, it is evident that model 5 is the best fit model.

Correlation:

	Adult obesity	*Adult smoking*	*Physical inactivity*	*Excessive drinking*	*Uninsured*	*Primary care physicians*	*High school graduation*	*Unemployment*	*Fast food restaurants*	*Dentists*
Adult obesity	1.00000									
Adult smoking	0.43014	1.00000								
Physical inactivity	0.78594	0.41153	1.00000							
Excessive drinking	-0.21619	0.46176	-0.23632	1.00000						
Uninsured	0.25600	0.02703	0.38659	-0.36361	1.00000					
Primary care physicians	0.71232	-0.00904	0.80238	-0.54897	0.45085	1.00000				
High school graduation	-0.18632	-0.03559	-0.45878	0.37642	-0.53652	-0.22592	1.00000			
Unemployment	0.51544	0.05711	0.68498	-0.46671	0.39717	0.90876	-0.15442	1.00000		
Fast food restaurants	0.56972	-0.23732	0.40290	-0.86279	0.32251	0.61701	-0.30026	0.43309	1.00000	
Dentists	0.82489	0.04857	0.77827	-0.47943	0.32516	0.91082	-0.25358	0.71342	0.63696	1.00000

According to Excel's Correlation Analysis, number if dentists has the highest correlation with obesity, at .82489. The number of people graduating high school has the lowest correlation with obesity, .18632.

Conclusion:

Below, is a summary of each model's results:

	Number of Variables	R Square	Adjusted R Square	F Ratio	Standard Error
Model 1	9	0.95191	0.80764	6.59820	0.01513
Model 2	8	0.9517	0.8551	9.8533	0.0131
Model 3	7	0.9382	0.8518	10.8527	0.0133
Model 4	6	0.9357	0.8714	14.5461	0.0124
Model 5	5	0.9329	0.8849	19.4567	0.0117
Model 6	4	0.9114	0.8672	20.5850	0.0126
Model 7	3	0.8532	0.8043	17.4378	0.0153

Based on this table, I conclude that Model 6, with five explanatory variables, is the best fit to the data. This model has the largest Adjusted R^2 and second largest F Ratio. It also has the smallest Standard Error. Although Model 5's R^2 is not the highest of group, the Adjusted R^2 is a better comparison since it adjusts for degrees of freedom. Additionally, the P-values of the explanatory variables in Model 5 are all fairly close to 0. For all these reasons, Model 5 is my choice for most precise model. This regression equation is:

$$Y = -0.1136 + 0.3353X_1 + 0.6848X_3 + 0X_5 - 0.2639X_7 + 0.3435X_8$$

Choosing Model 5 is saying that Adult smoking (%), Excessive drinking (%), Primary care physicians (#), Unemployment (%), and Fast food restaurants (%) are the main drivers of adult obesity in the selected counties of California.