LOGISTIC REGRESSION

By Daniel Lagos

Outside files used:

- The codebook for the 2009 California Health Interview Survey
- The dataset for the 2009 CHIS - named adult2009

1. STATE AN APPROPRIATE RESEARCH QUESTION that can be answered by multinomial or ordinal logistic regression. For example, "Can respondent health, rated from 1(=excellent) to 5(=poor) be predicted by socioeconomic status? (1 point).

   Research Question: Can health be predicted by poverty levels after adjusting for physical activity and depression?

   H_0: Health cannot be predicted by poverty levels after adjusting for physical activity and depression.

$$\boldsymbol{general\ health}(1 = \text{good or better}, 2 = \text{neutral}, 3 = \text{bad}) = \boldsymbol{poverty}(1 = 0 - 99\%\ \text{FPL}, 2 = 100 - 199\%\ \text{FPL}, 3 = 200 - 299\%\ \text{FPL}\ 4 = 300\%\ \text{FPL and above}) + \boldsymbol{depression}(1 = \text{no}, 2 = \text{yes}) + \boldsymbol{physical\ activity}(1 = \text{no/low}, 2 = \text{yes})$$

   FPL = Federal Poverty Level

2. DATA QUALITY ISSUES Select appropriate variables to answer the question posed. Conduct descriptive statistics to identify any data quality issues. (2 points).

   AB1: General Health Condition. Originally had 5 categories and was recoded as HEALTH (1 = good or better health, 2 = neutral health, 3 = bad health)

   POVLL: Poverty Level. Kept data and coding structure. Kept label. Coded as POVLL (1 = 0-99% FPL, 2= 100-199% FPL, 3 = 200-299% FPL, 4 = 300% FPL AND ABOVE)

   AJ32: Feel depressed in the last 30 days? Min = -2, which represents proxy missing. This value was restricted from the data set, leaving AJ32 with 5 levels. Changed label to DEPRESSED. Recoded as DEPRESSED (1 = not felt depressed in the last 30 days, 2 = have felt depressed in the last 30 days)

   AE25AMIN: How many minutes on average of doing vigorous activity per day? Continuous variable ranging from -1 to 480, where -1 = inapplicable. Restricted values of -1 from data set, leaving values equal to or over 0. Relabeled to PHYS and recoded as PHYS (1 = no/low, 2 = high), where "no/low" indicated 0 to 29 minutes of average vigorous activity per day. High indicated 30 minutes or more.

   Results presented as output of PROC MEANS procedure, note n = 12,944:

The MEANS Procedure

Variable	Label	N	N Miss	Minimum	Maximum	Mean	Std Dev	Kurtosis	Skewness
HEALTH	POVERTY LEVEL	12944	0	1.0000000	3.0000000	1.0995828	0.3443384	13.7236284	3.6713077
POVLL		12944	0	1.0000000	4.0000000	3.3746137	1.0020645	0.2938410	-1.3222537
DEPRESSED		12944	0	1.0000000	2.0000000	1.1248455	0.3305564	3.1542381	2.2701874
PHYS		12944	0	1.0000000	2.0000000	1.8132726	0.3897076	0.5856890	-1.6079797

3. DESCRIPTIVE ANALYSES: **Compute cross-tabulation with chi-square of categorical predictors**. **Compute ANOVA with numeric variables**. Don't have to do every variable. This is just exploratory analyses. (2 points).

Table 1

Table 1

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of DEPRESSED by POVLL

DEPRESSED	POVLL(POVERTY LEVEL) 1	2	3	4	Total
1	819 6.33 7.23 72.48	1268 9.80 11.19 79.65	1307 10.10 11.54 85.93	7934 61.29 70.04 91.18	11328 87.52
2	311 2.40 19.25 27.52	324 2.50 20.05 20.35	214 1.65 13.24 14.07	767 5.93 47.46 8.82	1616 12.48
Total	1130 8.73	1592 12.30	1521 11.75	8701 67.22	12944 100.00

Statistics for Table of DEPRESSED by POVLL

Statistic	DF	Value	Prob
Chi-Square	3	434.7829	<.0001
Likelihood Ratio Chi-Square	3	381.9223	<.0001
Mantel-Haenszel Chi-Square	1	433.2423	<.0001
Phi Coefficient		0.1833	
Contingency Coefficient		0.1803	
Cramer's V		0.1833	

Sample Size = 12944

We may have some differences when comparing DEPRESSED to POVLL. When depressed (DEPRESSED=2), we would expect to see near 12.5%. For levels of poverty 3 and below they are higher, especially with POVLL(1), POVLL(4) is under expected value of 12.5%. Chi-square results in p-value < 0.0001, well under alpha of 0.05. Statistically significant.

Sample Size = 12944

Frequency
Percent
Row Pct
Col Pct

Table of PHYS by POVLL

PHYS	POVLL(POVERTY LEVEL) 1	2	3	4	Total
1	269 2.08 11.13 23.81	389 3.01 16.09 24.43	338 2.61 13.98 22.22	1421 10.98 58.79 16.33	2417 18.67
2	861 6.65 8.18 76.19	1203 9.29 11.43 75.57	1183 9.14 11.24 77.78	7280 56.24 69.16 83.67	10527 81.33
Total	1130 8.73	1592 12.30	1521 11.75	8701 67.22	12944 100.00

Statistics for Table of PHYS by POVLL

Statistic	DF	Value	Prob
Chi-Square	3	98.4328	<.0001
Likelihood Ratio Chi-Square	3	95.3771	<.0001
Mantel-Haenszel Chi-Square	1	85.7653	<.0001
Phi Coefficient		0.0872	
Contingency Coefficient		0.0869	
Cramer's V		0.0872	

Sample Size = 12944

Again, possible differences when comparing PHYS to POVLL. When performing low levels of physical activity (PHYS=1), we would expect to see near 18.7%. For levels of poverty 3 and below they are higher. PHYS(4) is under expected value of 18.7%, but unsure how impactful this may be. Chi-square results in p-value < 0.0001, well under alpha of 0.05.

Statistically significant.

Table 2

The FREQ Procedure

Table of POVLL by HEALTH

POVLL(POVERTY LEVEL)	HEALTH 1	2	3	Total
1	859 6.64 76.02 7.25	222 1.72 19.65 24.26	49 0.38 4.34 26.20	1130 8.73
2	1300 10.04 81.66 10.98	240 1.85 15.08 26.23	52 0.40 3.27 27.81	1592 12.30
3	1373 10.61 90.27 11.59	120 0.93 7.89 13.11	28 0.22 1.84 14.97	1521 11.75
4	8310 64.20 95.51 70.17	333 2.57 3.83 36.39	58 0.45 0.67 31.02	8701 67.22
Total	11842 91.49	915 7.07	187 1.44	12944 100.00

Statistics for Table of POVLL by HEALTH

Statistic	DF	Value	Prob
Chi-Square	6	730.0984	<.0001
Likelihood Ratio Chi-Square	6	615.7069	<.0001
Mantel-Haenszel Chi-Square	1	672.3469	<.0001
Phi Coefficient		0.2375	
Contingency Coefficient		0.2311	
Cramer's V		0.1679	

Sample Size = 12944

Generally speaking definitely have differences in expected value in POVLL(4). The other levels may also have differences as well from expected value. Chi-square results in p-value < 0.0001, well under alpha of 0.05. Statistically significant.

Table of DEPRESSED by HEALTH

DEPRESSED	HEALTH 1	2	3	Total
1	10579 81.73 93.39 89.33	642 4.96 5.67 70.16	107 0.83 0.94 57.22	11328 87.52
2	1263 9.76 78.16 10.67	273 2.11 16.89 29.84	80 0.62 4.95 42.78	1616 12.48
Total	11842 91.49	915 7.07	187 1.44	12944 100.00

Statistics for Table of DEPRESSED by HEALTH

Statistic	DF	Value	Prob
Chi-Square	2	445.0989	<.0001
Likelihood Ratio Chi-Square	2	335.5537	<.0001
Mantel-Haenszel Chi-Square	1	441.4460	<.0001
Phi Coefficient		0.1854	
Contingency Coefficient		0.1823	
Cramer's V		0.1854	

Sample Size = 12944

Again, there are differences from expected value. The only debatable differences are all cases of DEPRESSED when compared to HEALTH(1). However, a logistic regression will provide conclusive results. Chi-square results in p-value < 0.0001, well under alpha of 0.05. Statistically significant.

Table of PHYS by HEALTH				
	HEALTH			
PHYS	1	2	3	Total
1	2120 16.38 87.71 17.90	248 1.92 10.26 27.10	49 0.38 2.03 26.20	2417 18.67
2	9722 75.11 92.35 82.10	667 5.15 6.34 72.90	138 1.07 1.31 73.80	10527 81.33
Total	11842 91.49	915 7.07	187 1.44	12944 100.00

Statistics for Table of PHYS by HEALTH			
Statistic	DF	Value	Prob
Chi-Square	2	54.4403	<.0001
Likelihood Ratio Chi-Square	2	49.8790	<.0001
Mantel-Haenszel Chi-Square	1	47.5819	<.0001
Phi Coefficient		0.0649	
Contingency Coefficient		0.0647	
Cramer's V		0.0649	

Sample Size = 12944

We have the same pattern as above, where all cases of PHYS when compared to HEALTH(1) may not be different from the expected value, but the other cases of PHYS compared to HEALTH(2) and HEALTH(3) may be impactful. Chi-square results in p-value < 0.0001, well under alpha of 0.05. Statistically significant.

Test for Multicollinearity

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	1.21988	0.02040	59.81	<.0001	0
POVLL	POVERTY LEVEL	1	-0.06796	0.00296	-22.93	<.0001	1.04073
DEPRESSED		1	0.15330	0.00896	17.11	<.0001	1.03502
PHYS		1	-0.03496	0.00750	-4.66	<.0001	1.00704

All Variance inflation factors are close to 1. We do not have a problem here.

Test for Ordinality

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
6.3592	5	0.2728

Proportional odds assumption is not violated based on failing to reject the null hypothesis that we have proportional odds (p-value = 0.2728). Must stay with ordinal logistic regression.

4. LOGISTIC REGRESSION Perform either a multinomial or ordinal logistic regression. (3 points).
    a. Using the dependent variable for ordinal logistic regression or multinomial logistic regression, compute a logistic regression.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	66.1	Somers' D	0.466
Percent Discordant	19.5	Gamma	0.545
Percent Tied	14.4	Tau-a	0.074
Pairs	13220989	c	0.733

C-value = 0.733. This is equivalent to AUC value in a ROC chart. Normally a c-value over 0.70 is considered favorable, 0.5 considered useless. The closer to 1 the better. We have a strong c-statistic. The accuracy of this model is strong.

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
POVLL 1 vs 4	1.0000	5.320	4.466	6.337
POVLL 2 vs 4	1.0000	4.085	3.462	4.821
POVLL 3 vs 4	1.0000	2.112	1.731	2.578
DEPRESSED 2 vs 1	1.0000	2.920	2.525	3.378
PHYS 1 vs 2	1.0000	1.451	1.252	1.682

Table 3. Adjusted ordinal logistic regression modeling for odds ratio of general health in CHIS 2009.

	General Health	
	OR†	95% CI†
Poverty		
300% FPL and above‡	--	--
200-299% FPL	2.112	(1.731, 2.578)
100-199% FPL	4.085	(3.462, 4.82)
0.00-99% FPL	5.320	(4.466, 6.337)
Depression		
No‡	--	--
Yes	2.920	(2.525, 3.378)
Physical Activity		
High‡	--	--
No/Low	1.451	(1.252, 1.682)

* Adjusted for all variables listed.
† OR = adjusted odds ratio; CI = 95% confidence interval.
‡ Reference category.

For every one unit decrease in health is associated with a 5.32-fold increase in odds for those in the 0.00-99% of Federal Poverty Levels after controlling for depression and physical activity (AOR=5.320; 95% CI = 4.466 to 6.337). For every one unit decrease in health is associated with a 4.085-fold increase in odds for those in the 100-199% Federal Poverty Levels after controlling for depression and physical activity (AOR=4.085; 95% CI = 3.462to 4.82). For every one unit decrease in health is associated with a 2.112-fold increase in odds for those in the 200-299% of

Federal Poverty Levels after controlling for depression and physical activity (AOR=2.112; 95% CI = 1.731 to 2.578).

For every one unit decrease in health is associated with a 2.920-fold increase in odds for those reporting feeling depressed within the last 30 days after controlling for poverty and physical activity (AOR=2.920; 95% CI = 2.525 to 3.378).

For every one unit decrease in health is associated with a 1.451-fold increase in odds for those reporting 0-29 minutes of vigorous physical activity after controlling for poverty and depression (AOR=1.451; 95% CI = 1.252 to 1.682).

b. Using the same dependent variable as in part a, compute a second logistic regression, with a subset of the predictor variables – **Removing Depression**.

First, we will update null hypotheses.

H_0: Health cannot be predicted by poverty levels after adjusting for physical activity.

Second, must provide results of Proportional Odds Assumption:

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
1.5123	4	0.8245

We are still in an ordinal logistic regression situation based on SAS output above. Still failing to reject null hypothesis.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	59.8	Somers' D	0.405
Percent Discordant	19.3	Gamma	0.512
Percent Tied	20.9	Tau-a	0.064
Pairs	13220989	c	0.703

C-statistic has dropped slightly, but still indicating strong accuracy at 0.703.

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
POVLL 1 vs 4	1.0000	6.531	5.509	7.742
POVLL 2 vs 4	1.0000	4.633	3.937	5.453
POVLL 3 vs 4	1.0000	2.244	1.842	2.735
PHYS 1 vs 2	1.0000	1.468	1.269	1.699

Table 3. Adjusted ordinal logistic regression modeling for odds ratio of general health in CHIS 2009.

	General Health	
	OR†	95% CI†
Poverty		
300% FPL and above‡	--	--
200-299% FPL	2.244	(1.842, 2.735)
100-199% FPL	4.633	(3.937, 5.453)
0.00-99% FPL	6.531	(5.509, 7.742)
Physical Activity		
High‡	--	--
No/Low	1.468	(1.269, 1.699)

* Adjusted for all variables listed.
† OR = adjusted odds ratio; CI = 95% confidence interval.
‡ Reference category.

For every one unit decrease in health is associated with a 6.531-fold increase in odds for those in the 0.00-99% of Federal Poverty Levels after controlling for physical activity (AOR=6.531; 95% CI = 5.509 to 7.742). For every one unit decrease in health is associated with a 4.633-fold increase in odds for those in the 100-199% Federal Poverty Levels after controlling for depression and physical activity (AOR=4.633; 95% CI = 3.937 to 5.453). For every one unit decrease in health is associated with a 2.244-fold increase in odds for those in the 200-299% of Federal Poverty Levels after controlling for depression and physical activity (AOR=2.244; 95% CI = 1.842 to 2.735).

For every one unit decrease in health is associated with a 1.468-fold increase in odds for those reporting 0-29 minutes of vigorous physical activity after controlling for poverty and depression (AOR=1.468; 95% CI = 1.269 to 1.699).

5. DISCUSSION Explain your results. State whether you reject or accept the entire model. **Interpret coefficients and odds ratios for each predictor**. State which of your two models is preferable. (2 points)

Model 1: HEALTH = f(POVERTY, PHYSICAL ACTIVITY, DEPRESSION)

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	828.1741	5	<.0001
Score	1009.9577	5	<.0001
Wald	829.3799	5	<.0001

Given that likelihood ratio chi-squared produces a p-value <0.0001, meaning we reject null hypothesis that Health cannot be predicted by poverty levels after adjusting for physical activity and depression, and C-value = 0.733 we can keep this model.

Model 2: HEALTH = f(POVERTY, PHYSICAL ACTIVITY)

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	640.2405	4	<.0001
Score	756.6507	4	<.0001
Wald	639.4437	4	<.0001

As before likelihood ratio chi-squared produces a p-value <0.0001, meaning we reject null hypothesis that Health cannot be predicted by poverty levels after adjusting for physical activity. In this case c-value has dropped slightly to 0.703.

Conclusion: Keep Model 1, as it has a higher c-value and model is statistically significant.