Daniel Prusinski

Take-home Final for 401

1. Your consulting group has been hired as political pollsters for the upcoming election. Your job is to estimate the proportion of voters who will vote for the Democratic ticket. You want to be 90% confident that your prediction is within 0.04 of the actual population proportion. (25 points total)

a. In order to achieve this, what sample size is needed?
.04 = 1.645 * Square Root of (.5*.5)/n
.04/1.645=.02431610942249240121580547112462
Square.024316 = .0005913
.5*.5=.25
.25/.0005913=422.79 Can not have .79 of a person so round up
423 people needed to be 90% confident that your prediction is within 0.04.
(5 points)

b. Suppose your fickle employers now want to have 95% confidence. What sample size will be needed? (5 points)
.04 = 1.96 * Square Root of (.5*.5)/n
.04/1.96=.020408
Square .020408 = .00041648
.5*.5=.25
.25/.00041648 = 600.26 Can not have .26 of a person so round up
601 People needed to be 95% confident that your prediction is within 0.04 of the actual population proportion.

c. They have changed their mind once more and now, along with 95% confidence, they want to know the sample size needed with a margin of error of 0.03. What sample size is needed? (5 points)
.03 = 1.96 * Square Root of (.5*.5)/n
.03/1.96 = .015306
Square .015306 = .000234277
.5*.5=.25
.25/.000234277 = 1067.11 Can not have .11 of a person so round up
1068 People needed to be 95% confident that your prediction is within 0.03 of the actual population proportion.

d. Based on your answers to these first three questions, what general conclusions can be drawn about how both the confidence level and margin of error affect the sample size needed, as well as how they interact with each other? (5 points)

The greater the confidence interval (CI) the greater the sample population is needed.
The smaller the margin of error the greater the sample population is needed.
Interacting together, the small the margin of error coupled with a larger CI interacts such that the sample population must be even larger. Also, if the CI were to lower and the margin of error were to increase the population needed would be smaller.

e. Assuming that it will cost your organization roughly $1 for each person you survey and you have a total of $250,000 to spend, what sampling plan in terms of confidence level, margin of error, and sample size would you recommend to management, and why? Please be as specific and quantitative as possible and be sure to justify your statements. (5 points)

In solving this problem, I am assuming that the $250,000 must all be spent on just the survey. Thus, I am asking myself, “Self, how can the campaign optimize its money to retain the most accurate measure from the survey?”

So, in my equation we will have a sample population of 250,000 people, thus utilizing all the money, as set forth by the first assumption.

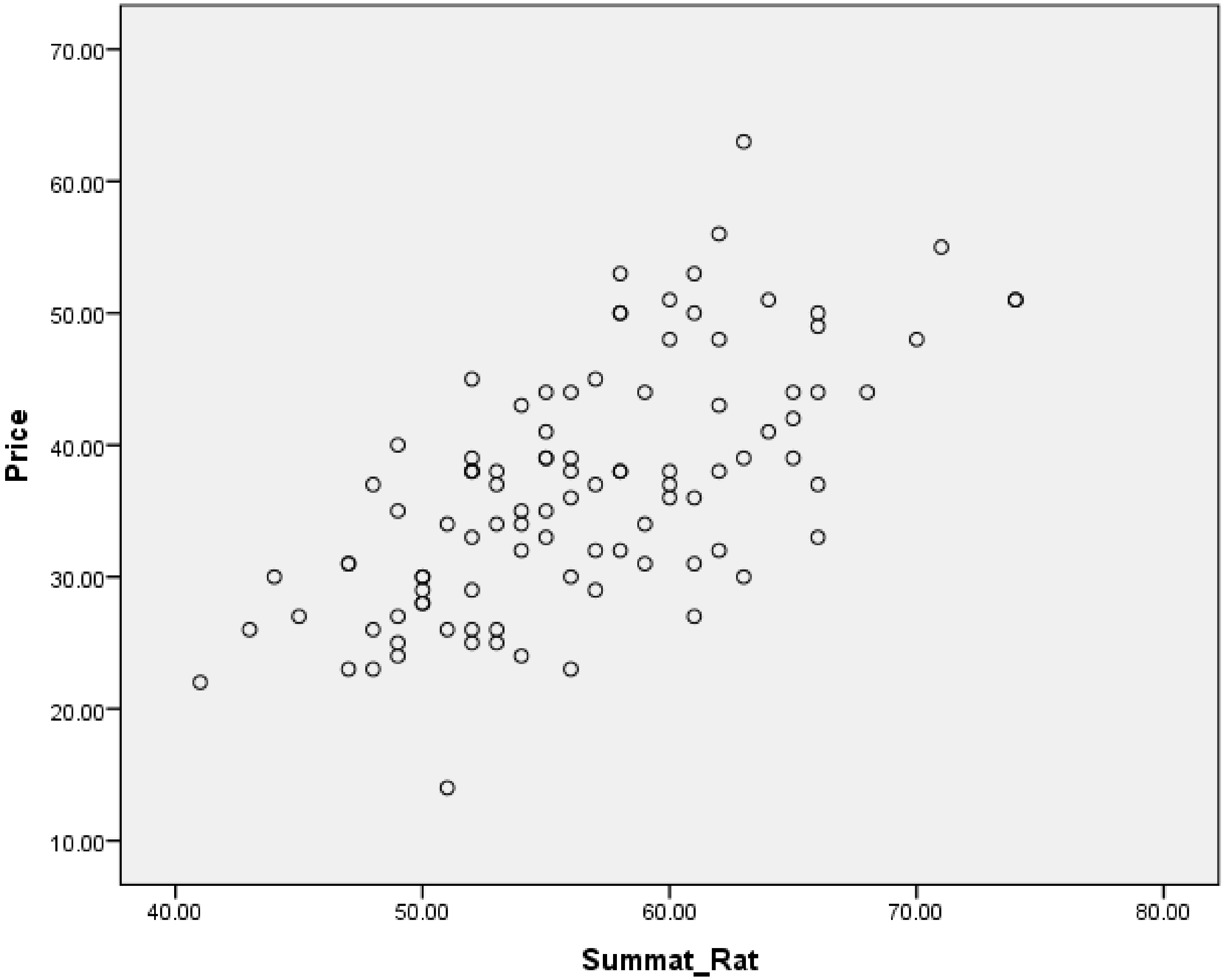
The confidence interval that makes the most sense is at the .001 level which gives us a great sense of confidence.

What I really value is the margin of error because this equates to specificity. When campaigning, candidates rely heavily on the leads or loss margins. This often dictates strategy and where they campaign. The mores accurate the margin of error, the more specific the candidate can focus the campaign machine. The campaign values knowing very accurately the lead or loss it currently has.

The $250,000 buys the survey a .00328 margin of error, which will buy them a great deal of specificity to discern lead or loss, with a 99.9 confidence interval which should help accurately focus on specific areas to campaign.

2. Zagat's publishes restaurant and hotel ratings for various establishments all over the world. The Excel file, Zagat Ratings, contains the ratings for food, décor, service, and price per person for a sampling of 50 restaurants located in New York City and 50 located in Long Island. Your group has been asked by Mr. Zagat himself to develop a regression model to predict the average price per person based on the sum of the ratings for food, décor and service. For this question, you must use SPSS or Excel for the relevant parts. Please use the attached data set. (45 points total)

a. Set up a scatter diagram with the summated ratings on the horizontal axis and price per person on the vertical axis. (5 points)



b. Assuming a linear relationship, use the least-squares method to find the regression equation and the coefficients b_0 (intercept) and b_1 (slope) and interpret their meaning. (5 points)

Y = -13.656 +.893x

Intercept: The slope tells us that for every additional rating point the price increases .893 dollars.

At a 0 rating one would have a bill of -13.656, but there is no such point. This is helpful for determining the 0 point on the y axis.

c. Use the regression equation from part b to predict the average price of a restaurant with a summated rating of 50. (5 points)

-13.656 + .893*50 = $30.99

d. The predicted value *y-hat* in part c) is an estimated average price per person. The quantity SSE (Sum of Squared Errors) divided by *n-2* gives you an estimate of the variance ʃ2 of the error term. Using this information, compute a standard error and a 95% confidence interval (NOT prediction interval) for the actual average price per person for a restaurant with a summated rating of 50. (5 points)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3560.349	1	3560.349	72.315	.000[b]
	Residual	4824.891	98	49.234		
	Total	8385.240	99			

a. Dependent Variable: Price

b. Predictors: (Constant), Sum_Rating

In this problem, the standard error is 4824.891, and n= 100. 4824.891/98= 49.23

At $50 the actual standard error is .44721 and at the 95% confidence interval the range is 27.7583 and 30.2417.

Descriptives

Price

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
41.00	1	22.0000	.	.	.	.	22.00	22.00
43.00	1	26.0000	.	.	.	.	26.00	26.00
44.00	1	30.0000	.	.	.	.	30.00	30.00
45.00	1	27.0000	.	.	.	.	27.00	27.00
47.00	3	28.3333	4.61880	2.66667	16.8596	39.8071	23.00	31.00
48.00	3	28.6667	7.37111	4.25572	10.3558	46.9775	23.00	37.00
49.00	5	30.2000	6.97854	3.12090	21.5350	38.8650	24.00	40.00
50.00	5	29.0000	1.00000	.44721	27.7583	30.2417	28.00	30.00
51.00	3	24.6667	10.06645	5.81187	-.3398	49.6731	14.00	34.00
52.00	9	34.5556	6.72888	2.24296	29.3833	39.7278	25.00	45.00
53.00	5	32.0000	6.12372	2.73861	24.3964	39.6036	25.00	38.00
54.00	5	33.6000	6.80441	3.04302	25.1512	42.0488	24.00	43.00
55.00	6	38.5000	3.98748	1.62788	34.3154	42.6846	33.00	44.00
56.00	6	35.0000	7.42967	3.03315	27.2030	42.7970	23.00	44.00
57.00	4	35.7500	6.99405	3.49702	24.6209	46.8791	29.00	45.00
58.00	6	43.5000	8.57321	3.50000	34.5030	52.4970	32.00	53.00
59.00	3	36.3333	6.80686	3.92994	19.4242	53.2425	31.00	44.00
60.00	5	42.0000	6.96419	3.11448	33.3528	50.6472	36.00	51.00
61.00	5	39.4000	11.54556	5.16333	25.0643	53.7357	27.00	53.00
62.00	5	43.4000	9.20869	4.11825	31.9659	54.8341	32.00	56.00
63.00	3	44.0000	17.05872	9.84886	1.6238	86.3762	30.00	63.00
64.00	2	46.0000	7.07107	5.00000	-17.5310	109.5310	41.00	51.00
65.00	3	41.6667	2.51661	1.45297	35.4151	47.9183	39.00	44.00
66.00	5	42.6000	7.43640	3.32566	33.3665	51.8335	33.00	50.00
68.00	1	44.0000	.	.	.	.	44.00	44.00
70.00	1	48.0000	.	.	.	.	48.00	48.00
71.00	1	55.0000	.	.	.	.	55.00	55.00
74.00	2	51.0000	.00000	.00000	51.0000	51.0000	51.00	51.00
Total	100	36.7400	9.20323	.92032	34.9139	38.5661	14.00	63.00

At $50 the actual standard error is .44721 and at the 95% confidence interval the range is 27.7583 and 30.2417.

e. Remember that the computed value of the slope is a sample-based estimate of the unknown actual slope. Compute a 95% confidence interval estimate of the actual slope, . (5 points)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-13.656	5.968		-2.288	.024	-25.499	-1.813
	Sum_Rating	.893	.105	.652	8.504	.000	.685	1.102

a. Dependent Variable: Price

The 95% confidence interval estimate of the actual slope B is between .685 and 1.102.

f. Compute the coefficient of determination and interpret its meaning. (5 points)

Model Summary[b]

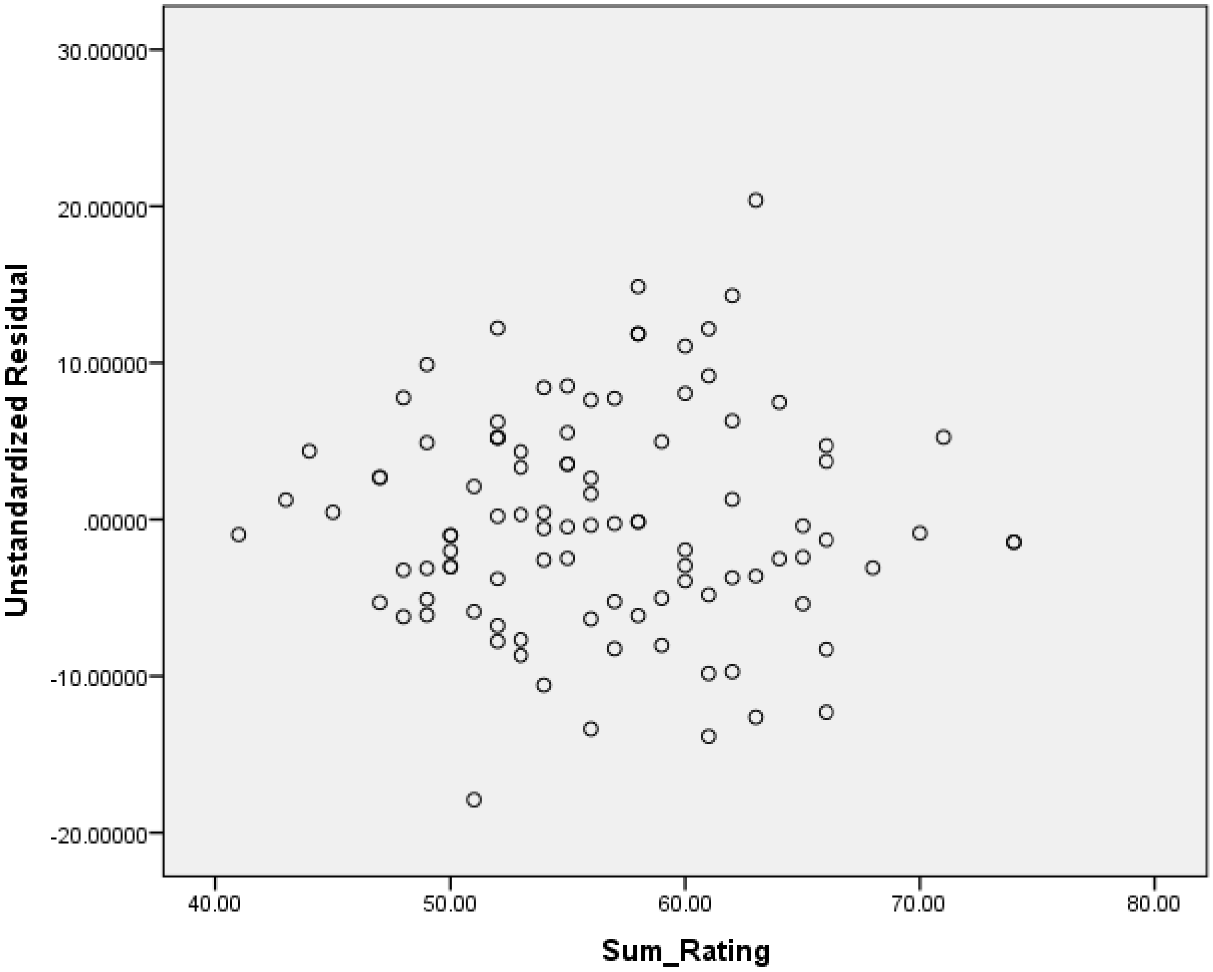
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.652[a]	.425	.419	7.01666

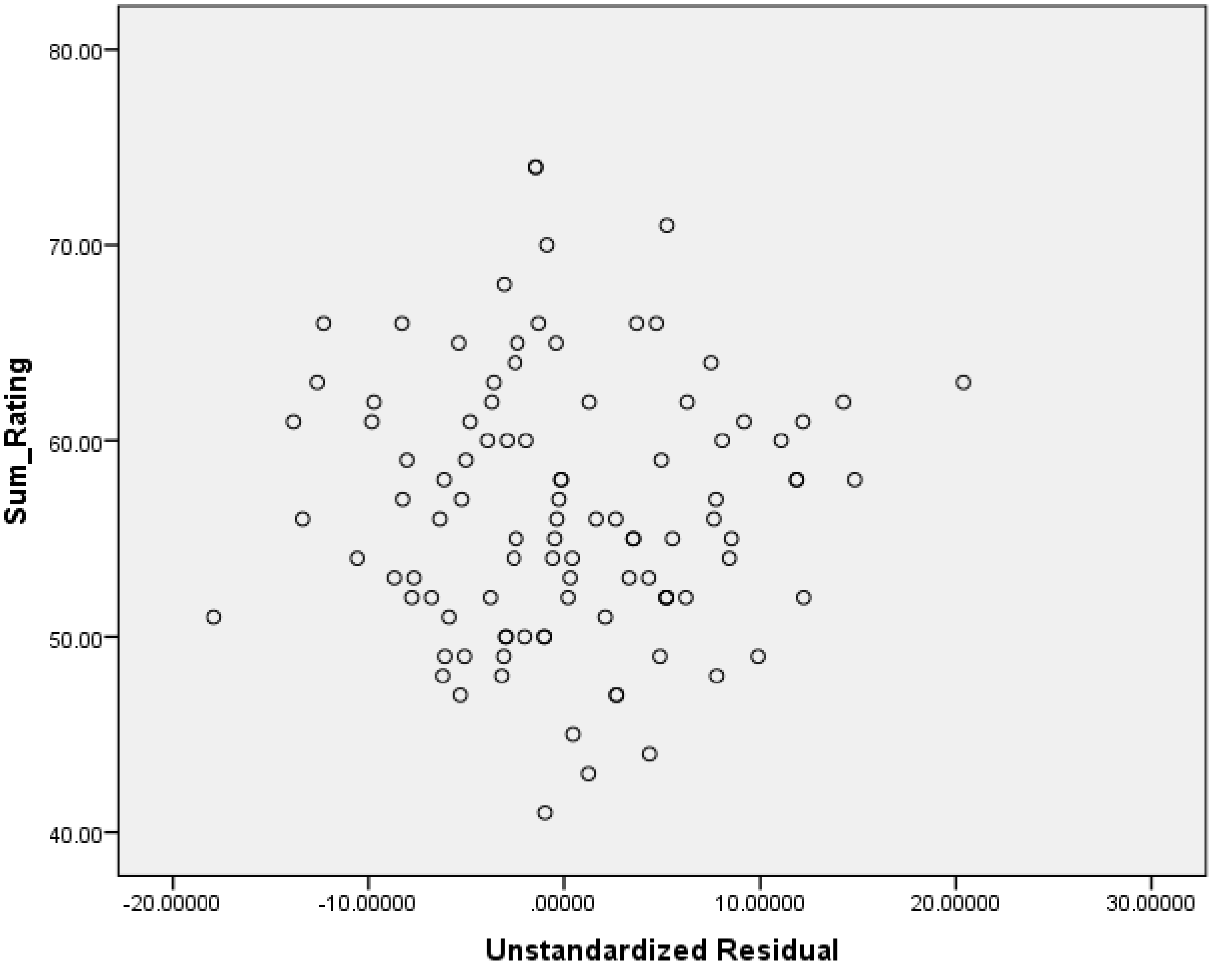
a. Predictors: (Constant), Sum_Rating

b. Dependent Variable: Price

The coefficient of determination is .425, and 42.5% percent of the price is determined by the summated rating. The closer to 1 the coefficient rating is the stronger its ability to explain price. In this case, .425 is not a strong score but it is also not a weak score either. In looking at the data, I would interpret this score to be helpful but not a definitive predictor of price. The next step I would do would be to see if I can't add another independent variable to strengthen my overall Rsquared.to better predict price.

g. Perform a simple residual analysis on your results and determine the adequacy of the fit of the model. You will need to plot the errors (difference between actual and predicted *y* values) and determine to what extent they are independent of the *x*-values. (5 points)





The goal in a residual plot is to see a relatively random plot of residuals. In this case, there is a relatively random dispersion, which makes the plots relatively unbiased of the x-values. The assumptions for multiple regression are met given that the residual plot is random and there are not any extreme outliers making a bias in the regression analysis. The residuals are somewhat independent, which validates the rsquared, but still validate the overall regression.

h. At the 0.05 level of significance, is there evidence of a linear relationship between the price per person and the summated ratings? (5 points)

Correlations

		Sum_Rating	Price
Sum_Rating	Pearson Correlation	1	.652**
	Sig. (2-tailed)		.000
	N	100	100
Price	Pearson Correlation	.652**	1
	Sig. (2-tailed)	.000	
	N	100	100

**. Correlation is significant at the 0.01 level (2-tailed).

The estimated linear relationship or correlation coefficient is .652, and this demonstrates a positive linear relationship between summated rating and price. The relationship is different from zero, because p<.001 which is beyond the .05 significance level.

i. How useful do you think the summated rating is as a predictor of price per person?
Explain. Given the information communicated in the problem, I find that the summated rating is a relatively accurate predictor of price. This problem has a strong r and Rsquared, there are quite a few data points to draw from, and the confidence interval is strong. As stated earlier, I would want to see if I could add other independent variables to strengthen my Rsquared, which would strengthen my dependent variable of price.
(5 points)

3. Your company is considering organizational changes based on adopting the use of self-managed work teams. To assess the attitudes of employees of the company toward this change, a sample of 400 employees is selected and asked whether they favor the introduction of self-managed work teams in the organization. Three responses were permitted: like, indifferent, or despise. The results are in the table below, classified by job. (30 points total)

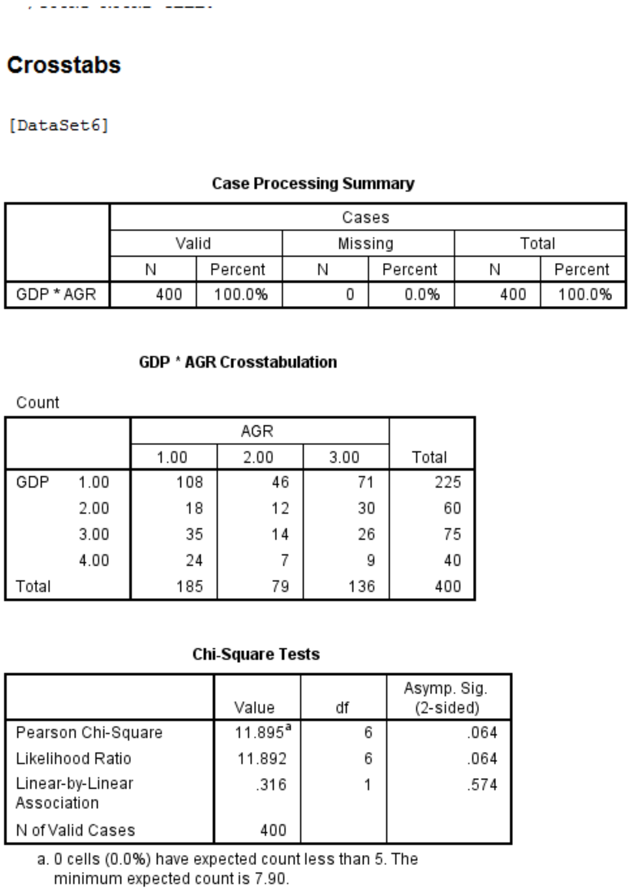
Attitudes toward self-managed work teams

Job Type Like Indifferent Despise Total

Hourly
Worker
108 46 71 225
Supervisor 18 12 30 60
Middle
Management
35 14 26 75
Upper

Management
24 7 9 40
Total 185 79 136 400

a. At the 0.05 level of significance, is there evidence of a relationship between attitude toward self-managed work teams and type of job? Use the chi-squared method to answer this question, and show your calculations. (10 points)



Crosstabs

[DataSet6]

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
GDP * AGR	400	100.0%	0	0.0%	400	100.0%

GDP * AGR Crosstabulation

Count

		AGR			Total
		1.00	2.00	3.00	
GDP	1.00	108	46	71	225
	2.00	18	12	30	60
	3.00	35	14	26	75
	4.00	24	7	9	40
Total		185	79	136	400

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	11.895[a]	6	.064
Likelihood Ratio	11.892	6	.064
Linear-by-Linear Association	.316	1	.574
N of Valid Cases	400		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 7.90.

At the 0.05 level of significance there is not evidence of a relationship between attitude toward self-managed work teams and type of job because the value is 11.895 and at .05 it needs to be 12.59.

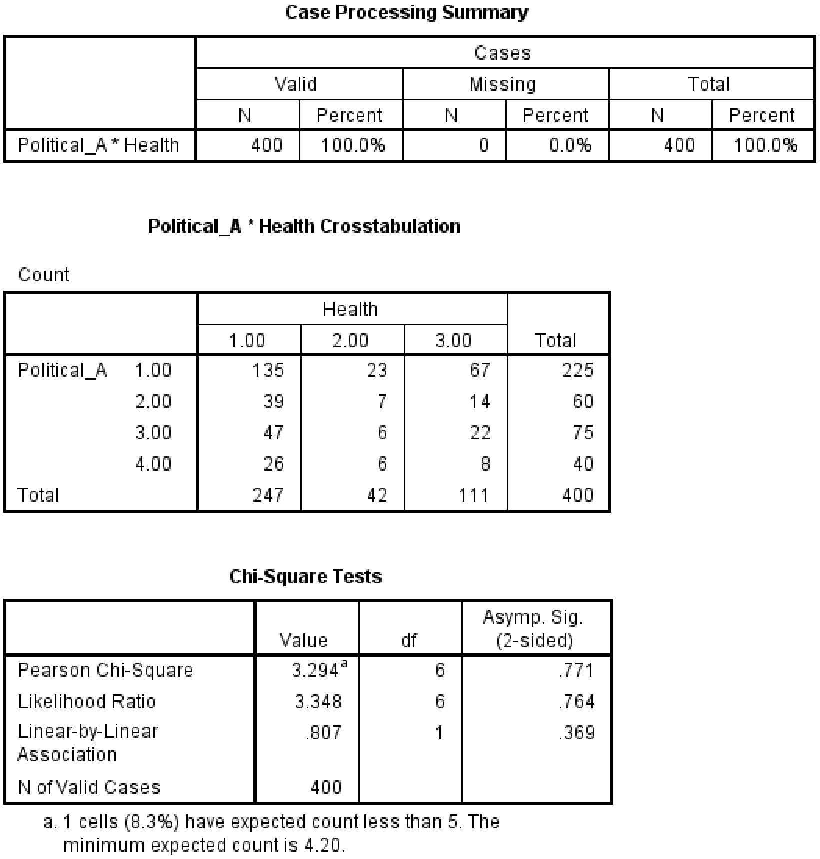
That being said, at the .10 level there is a statistical significant relationship, this is important for analyzing this data with the next question. Also, as someone interpreting this data, I would let management know that there appears to be some relationship, but it is not at the .05 level.

The survey went on to ask respondents about their attitudes toward instituting a policy whereby an employee could take one additional vacation day per month without pay. The results are in the table below.

Attitudes toward vacation sans pay

Job Type Like Indifferent Despise Total
Hourly
Worker
135 23 67 225
Supervisor 39 7 14 60
Middle
Management
47 6 22 75
Upper
Management
26 6 8 40
Total 400

b. At the 0.05 level of significance, is there any evidence of a relationship between attitude toward vacation time without pay and type of job? Use the chi-squared method to answer this question, and show your calculations. (10 points)



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Political_A * Health	400	100.0%	0	0.0%	400	100.0%

Political_A * Health Crosstabulation

Count

		Health			Total
		1.00	2.00	3.00	
Political_A	1.00	135	23	67	225
	2.00	39	7	14	60
	3.00	47	6	22	75
	4.00	26	6	8	40
Total		247	42	111	400

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	3.294[a]	6	.771
Likelihood Ratio	3.348	6	.764
Linear-by-Linear Association	.807	1	.369
N of Valid Cases	400		

a. 1 cells (8.3%) have expected count less than 5. The minimum expected count is 4.20.

At the 0.05 level of significance there is not evidence of a relationship between attitude toward self-managed work teams and type of job because the value is 3.294 and at .05 level of significance it needs to be 12.59. Furthermore, if I were reporting this to management I would highlight that the Pearson Chi-Square is significantly lower than the needed score to show significance, thus further demonstrating that there is not a relationship.

c. Explain how the answers to parts a and b may or may not be related. Also explain one other measure of association that you could use to analyze these data and answer the questions. It is not necessary to re-solve the questions in entirety from scratch with your alternate method. Just explain the procedure. (10 points)sdf

From visually analyzing the data it would appear the there is an association between “like” for Hourly Workers, Middle Management, and Upper management for both surveys. There also seems to be a relationship in polarity of response for both surveys. Either workers liked an idea or despised it. It might appear that those that like self-management might also like taking one day of no pay per month for vacations purposes. These are loose associations that one gathers from looking at both surveys, but it does not prove a causation relationship.

Given that the data can be organized with one side being ordinal and the other being nominal, one could use the Lambda approach.- Seeing that in the above problems we can not find a dependent/independent relationship, we would have to use the Lambda Symmetric approach. Using this approach one would put everyone in the largest category and keep track of the number of error of assignments made. This information would be expressed as a percentage or proportion of the entire population. This would be done for both the columns and rows. After each proportion was computed, we would average the two proportions which would yield the Lambda symmetric approach. Given that the Chi-square did not prove relationships at a 95% confidence interval, I would not expect a clearer depiction of relationship with the Lambda approach, but it would be helpful for somewhat validating results.