

STA 303 H1S / 1002 HS – Winter 2011
Test
March 7, 2011

LAST NAME: _____ FIRST NAME: _____

STUDENT NUMBER: _____

ENROLLED IN: (circle one) STA 303 STA 1002

INSTRUCTIONS:

- Time: 90 minutes
- Aids allowed: calculator.
- Some formulae are on the last page (page 11).
- Total points: 45

1ab	1cde	2abcde	2fghij	3

1. Many countries survey a random sample of adults every year or two to collect demographic information and opinions on issues ranging from government spending to the state of race relations to the existence and nature of God.

In this question, we will consider data collected from 943 people aged 18-30 in the 2004 American General Social Survey. The variables we will consider are gender (male or female), political party affiliation (Democrat, Independent, or Republican), and response on a 7 point scale to a question rating their political ideology where 1=extremely liberal, 2=liberal, 3=slightly liberal, 4=moderate, 5=slightly conservative, 6=conservative, and 7=extremely conservative.

For this question, our interest is in how the levels of liberalism or conservatism in the responses to the political ideology question vary with political party affiliation and gender. Thus we will treat political ideology level as a quantitative variable (variable name: **ideology**) and examine how its mean differs among the groups of subjects, categorized by political party affiliation (variable name: **party**) and gender (variable name: **gender**). The higher the mean ideology, the more conservative the group's responses tended to be.

Some edited SAS output from 2 models is below and on the next page. Some numbers have been replaced by X's.

MODEL 1

The GLM Procedure

Class Level Information

Class	Levels	Values
party	3	Democrat Independent Republican
gender	2	Female Male

Number of Observations Read	943
Number of Observations Used	943

Dependent Variable: ideology

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	90.332492	18.066498	10.81	<.0001
Error	937	1565.885959	1.671170		
Corrected Total	942	1656.218452			

R-Square	Coeff Var	Root MSE	ideology Mean
0.054541	31.55711	1.292737	4.096501

Source	DF	Type III SS	Mean Square	F Value	Pr > F
party	2	87.79540434	43.89770217	26.27	<.0001
gender	1	1.48829759	1.48829759	0.89	0.3456
party*gender	X	XXXXXXXXXX	XXXXXXXXXX	XXXX	0.3370

(SAS output for question 1 continues on the next page)

(SAS output for question 1 continued)

Parameter		Estimate	Standard Error	t Value	Pr > t
Intercept		4.664062500 B	0.11426291	40.82	<.0001
party	Democrat	-0.896062500 B	0.16255882	-5.51	<.0001
party	Independent	-0.622395833 B	0.15703935	-3.96	<.0001
party	Republican	0.000000000 B	.	.	.
gender	Female	-0.231963735 B	0.15287863	-1.52	0.1295
gender	Male	0.000000000 B	.	.	.
party*gender	Democrat Female	0.310475362 B	0.21098368	1.47	0.1415
party*gender	Democrat Male	0.000000000 B	.	.	.
party*gender	Independent Female	0.142959790 B	0.21181548	0.67	0.4999
party*gender	Independent Male	0.000000000 B	.	.	.
party*gender	Republican Female	0.000000000 B	.	.	.
party*gender	Republican Male	0.000000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

MODEL 2

(Initial output that is the same as for MODEL 1 has been deleted)

Dependent Variable: ideology

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	86.692970	28.897657	17.29	<.0001
Error	939	1569.525482	1.671486		
Corrected Total	942	1656.218452			

R-Square	Coeff Var	Root MSE	ideology Mean
0.052344	31.56010	1.292860	4.096501

Source	DF	Type III SS	Mean Square	F Value	Pr > F
party	2	84.25160525	42.12580263	25.20	<.0001
gender	1	1.31102845	1.31102845	0.78	0.3760

Parameter		Estimate	Standard Error	t Value	Pr > t
Intercept		4.576815040 B	0.08971327	51.02	<.0001
party	Democrat	-0.711248269 B	0.10353591	-6.87	<.0001
party	Independent	-0.542288453 B	0.10538744	-5.15	<.0001
party	Republican	0.000000000 B	.	.	.
gender	Female	-0.075780010 B	0.08556576	-0.89	0.3760
gender	Male	0.000000000 B	.	.	.

(Questions related to this output begin on the next page.)

(Question 1 continued)

- (a) (4 marks) Write the **model** that is being estimated in the output labelled MODEL 1; define all variables.

- (b) For the test in MODEL 1 with p -value 0.3370:

i. (1 mark) What are the null and alternative hypotheses?

ii. (2 marks) Explain in *practical* terms what you conclude from the test.

iii. (4 marks) What are the 4 missing numbers?

DF = _____

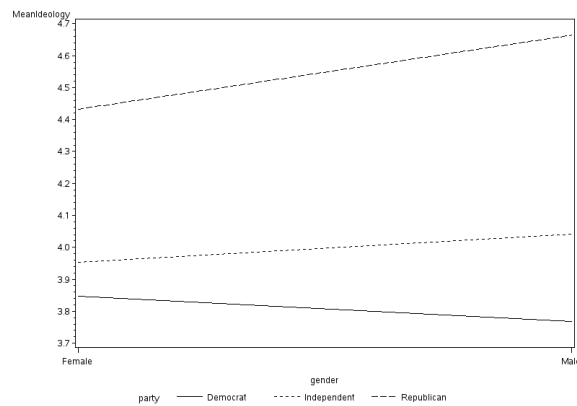
Type III SS = _____

Mean Square = _____

F Value = _____

(Question 1 continued)

- (c) (1 mark) For MODEL 2, what practical quantity, if anything, is being estimated by the estimate of the intercept?
- (d) (1 mark) For MODEL 2, estimate the mean political ideology score for females whose party affiliation is Democrat.
- (e) (6 marks) Here is a plot showing the mean value of political ideology for each gender, with separate lines for each party affiliation. (Females are on the left and males are on the right. The top line is for Republicans, middle line is for Independents, and bottom line is for Democrats.)

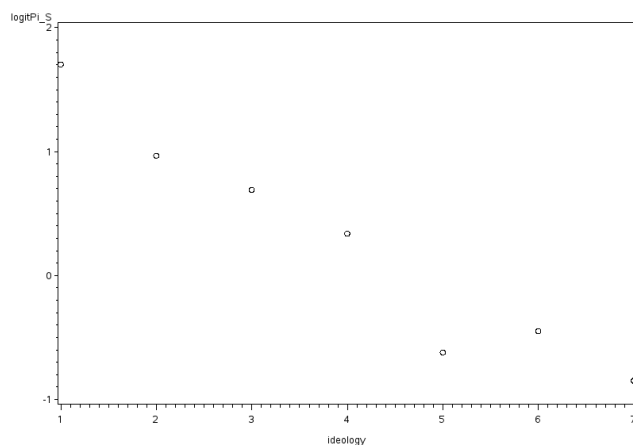


Explain how the interaction plot is consistent with the conclusions that can be drawn from inferences about the fitted model. Support your answer with relevant numbers from the SAS output.

2. In this question, we will work with the same data as in question 1. However, we will only consider people whose party affiliation is Democrat or Republican (people who identified themselves as Independent have been removed from the data) and we will ignore gender. Here we will consider a model for how well party affiliation can be predicted from a person's political ideology score.

Of the m_i people who responded that their political ideology is level i ($i = 1, \dots, 7$), y_i (variable name: `nDemocrats`) is the number whose party affiliation is Democrat. Political ideology (variable name: `ideology`) is treated as a quantitative explanatory variable in this analysis.

Some SAS output is given below and on the next page. The plot is the logit of the response proportions versus political ideology score. A few numbers in the output have been replaced by X's.



The LOGISTIC Procedure

Model Information

Data Set	WORK.BINOMIAL
Response Variable (Events)	nDemocrats
Response Variable (Trials)	m
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	7
Number of Observations Used	7
Sum of Frequencies Read	630
Sum of Frequencies Used	630

Response Profile

Ordered Value	Binary Outcome	Total Frequency
1	Event	340
2	Nonevent	290

(SAS output for this question continues on the next page.)

(SAS output for question 2 continued)

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	6.0544	5	1.2109	0.3010
Pearson	5.9719	5	1.1944	0.3090

Number of events/trials observations: 7

Model Fit Statistics

Criterion	Intercept	Intercept and Covariates
	Only	
AIC	871.393	827.862
SC	875.839	836.753
-2 Log L	869.393	823.862

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	XXXXXXX	1	<.0001
Score	XXXXXXX	1	<.0001
Wald	XXXXXXX	1	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard	Wald	Pr > ChiSq
			Error	Chi-Square	
Intercept	1	1.9110	0.2864	44.5265	<.0001
ideology	1	-0.4194	0.0652	41.3812	<.0001

Odds Ratio Estimates

Effect	Point	95% Wald	
	Estimate	Confidence Limits	
ideology	XXXXX	XXXXX	XXXXX

Obs	ideology	m	n	Pearson	DevRes
			Democrats	Res	
1	1	13	11	0.27773	0.28405
2	2	69	50	-0.38861	-0.38534
3	3	90	60	0.18010	0.18049
4	4	238	139	0.80536	0.80719
5	5	100	35	-2.08213	-2.10377
6	6	100	39	0.77133	0.76572
7	7	20	6	0.36400	0.35908

(Questions related to this output begin on the next page.)

(Question 2 continued)

- (a) (2 marks) On page 6 you are given a plot of the logit of the response proportions versus political ideology score. What is the purpose of looking at this plot and what do you conclude from it?
- (b) (2 marks) Why is the “Number of Observations Read” 7? And how is the “Sum of Frequencies Read” arrived at?
- (c) (3 marks) Write the log-likelihood function in terms of y_i , m_i , and the model parameters to be estimated.
- (d) (2 marks) What is the fitted equation? Define all variables.
- (e) (1 mark) How would your answer to part (d) change if y_i was changed to the number of people whose party affiliation was Republican?

(Question 2 continued)

- (f) (1 mark) From the model, what is the estimate of the probability of belonging to the Democratic party for someone whose response to the political ideology question is 4?
- (g) (3 marks) What is the odds ratio estimate for the effect of ideology? For higher responses on the political ideology scale, are the odds of being a Democrat higher or lower? By how much?
- (h) (1 mark) Write down the saturated model.
- (i) (4 marks) From the SAS output that you are given, can you carry out a Likelihood Ratio Test to test whether the coefficient of ideology is statistically significantly different from 0? If yes, carry it out, giving each of the following: (I) the test statistic, (II) the distribution of the test statistic under the null hypothesis, (III) the p -value, (IV) the conclusion. If no, state what is needed to find each of these 4 things.
- (j) (1 mark) You are given the Pearson (**PearsonRes**) and Deviance (**DevRes**) residuals? What can you conclude from them?

3. (6 marks) Fill out the table below to compare and contrast features of the models used in questions 1 and 2.

	First Model in Question 1	Model in Question 2
Underlying probability distribution of the response (You do not need to specify the parameters of the distribution.)		
Condition that must hold regarding the variance in order for inferences to be valid		
Probability distribution used to calculate the p -value for the test with null hypothesis that the coefficients of all parameters except the intercept are 0 (You do not need to specify the parameters (or df) of the distribution.)		

Some formulae:

Pooled t -test

$$t_{obs} = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Linear Regression

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \quad b_0 = \bar{y} - b_1 \bar{x}$$

One-way analysis of variance

$$\text{SSTO} = \sum_{i=1}^N (y_i - \bar{y})^2 \quad \text{SSE} = \sum_{g=1}^G \sum_{(g)} (y_i - \bar{y}_g)^2$$

$$\text{SSR} = \sum_{g=1}^G n_g (\bar{y}_g - \bar{y})^2$$

Bernoulli and Binomial distributions

$$\begin{aligned} &\text{If } Y \sim \text{Bernoulli}(\pi) \\ &\text{E}(Y) = \pi, \text{Var}(Y) = \pi(1 - \pi) \end{aligned}$$

$$\begin{aligned} &\text{If } Y \sim \text{Binomial}(m, \pi) \\ &\text{E}(Y) = m\pi, \text{Var}(Y) = m\pi(1 - \pi) \end{aligned}$$

Logistic Regression with Binomial Response formulae

$$\text{Deviance} = 2 \sum_{i=1}^n \{y_i \log(y_i) + (m_i - y_i) \log(m_i - y_i) - y_i \log(\hat{y}_i) + (m_i - y_i) \log(m_i - \hat{y}_i)\}$$

$$D_{res,i} = \text{sign}(y_i - m_i \hat{\pi}_i) \sqrt{2 \left\{ y_i \log \left(\frac{y_i}{m_i \hat{\pi}_i} \right) + (m_i - y_i) \log \left(\frac{m_i - y_i}{m_i - m_i \hat{\pi}_i} \right) \right\}}$$

$$P_{res,i} = \frac{y_i - m_i \hat{\pi}_i}{\sqrt{m_i \hat{\pi}_i (1 - \hat{\pi}_i)}}$$

Model Fitting Criteria

$$\text{AIC} = -2 \log(L) + 2(p + 1)$$

$$\text{SC} = -2 \log(L) + (p + 1) \log(N)$$