Introduction to Econometric Methods Final Exam
2007.9.14

There are four parts; all points are assigned in parentheses. Please write down the calculation process of your answers for partial credits will be given to you.

Part A:

Investigate the following table, and answer the following questions:

2000 Presidential Election Voting Choice by Ethnicities

	Y	**1**	**2**	**3**	**Total Count**
		Voted for			
X		**Soong**	**Lien**	**Chen**	
1	**Mainlanders**	98	29	22	149
2	**Hakha**	47	33	38	118
3	**Holo**	224	183	383	791
Total Count		369	245	443	1,058

Source: Telephone survey results, Election Study Center, NCCU, 2000.3.

1. (5%) Suppose frequencies represent probabilities. Fill in the joint and marginal probabilities of each cell below:

	Y	**1**	**2**	**3**	**Total**
		Voted for			
X		**Soong**	**Lien**	**Chen**	
1	**Mainlanders**				
2	**Hakha**				
3	**Holo**				
Total					1.000

2. (5%) Calculate E(Y | X = 1), E(Y | X = 2) and E(Y | X = 3).
3. (5%) A randomly selected voter of this pool reports voting for Mr. Chen. What is the probability that this voter belongs to the ethnicity group of Mainlanders? Of Hakha? Of Holo?
4. (5%) Are voting choice and ethnicity group membership independent? Explain.
5. (5%) Suppose standard deviations can serve as an indicator of diversity. Which candidate has the most diverse supporter base? Which has the least diverse base?
6. (5%) Which ethnicity group has the most diverse voting choices? Which has the least diverse choices?
7. (Bonus: 5%) What does this table tell us about ethnicity-based politics?

Part B:

A website http://www.columbia.edu/itc/ealac/sobelman/yic/taiwan_daxuan/article_taiwan_election.htm at Columbia University recorded the following story: During Jan. 15-17, 2004, the United Daily News (UDN) conducted a telephone survey of 1767 likely voters for the upcoming Taiwanese presidential election. Among all 1,767 subjects, 699 of them declined to answer, making the successful sample size 1,068. Among the successful samples, 374 responded that they would vote for the incumbents (Chen-Lu) and 449 responded that they would vote for the challengers (Lien-Soong). The remaining 245 voters claimed to be undecided. Let p denote the fraction of all likely voters that preferred the challengers at the time of the survey, and let $\hat{p}$ be the fraction of survey respondents that preferred the challengers.

1. (5%) Use the survey results to estimate p assuming all undecided voters would vote like the decided ones.
2. (5%) Drop all the undecided voters, and use the survey results to estimate p. How is this different from 1.?

For the following questions, simply ignore all undecided voters and act as if there were 823 voters surveyed, in which 374 responded that they would vote for the incumbents and 449 responded that they would vote for the challengers.

3. (5%) Use the estimator of the variance of $\hat{p}$, $\hat{p}(1-\hat{p})/n$, to calculate the standard error of your estimator.
4. (5%) What is the p-value for the test H_0: $p = 0.5$ vs. H_1: $p \neq 0.5$? What is the p-value for the test H_0: $p = 0.5$ vs. H_1: $p > 0.5$? Why do the two results differ?
5. (5%) Did the survey contain statistically significant evidence that the challengers were ahead of the incumbents at the time of the survey? Explain.
6. (Bonus: 5%) If the actual results of the election do not confirm the prediction of the survey, what are the possible reasons other than "bad luck"? One obvious reason is that something else happened after the date of the survey. Please list possible reasons OTHER than that.
7. (Bonus: 5%) The actual election had the incumbents and challengers split nearly at 50-50, so UDN's prediction was way off. Which of the reason you listed in 6. do you think contributed to UDN's miss?

Part C:

The latest surveying statistics show that among the universities in Taiwan, 32 public universities have a sample average (main) campus (land) size ($\overline{Y}$) of 644,864m^2, and a sample standard deviation (s_Y) of 581,618 m^2.

1. (5%) You would like to describe to your American friends the distribution of campus size for Taiwanese public universities using this sample average and standard deviation. However, your friend only understands land size in square

feet. Knowing that 1ft = 0.3048m, what would you tell your friend?

2. (5%) Construct the 95% confidence interval for the mean campus size of all public universities in Taiwan.
3. (5%) Suppose the same survey shows that 30 private universities have a sample average campus size of 283,557 m^2, and a sample standard deviation of 238,078 m^2. Construct a 90% confidence interval for the difference in mean campus sizes between public and private universities in Taiwan.
4. (5%) Can you conclude with a high degree of confidence that the population means for public and private universities are different? (What is the standard error of the difference in the two sample means? What is the *p*-value of the test of no difference in means versus some difference?)
5. (Bonus: 5%) What are the possible differences between public and private schools in Taiwan that may contribute to this difference in campus size? Explain.

Part D: [Not-so-simple Regression]

A data analyst wishes to estimate the relationship between *X*, *Z* and *Y*. She has *n* observation of each variable. Define the vectors $X^0 = (1,\ldots,1)'$, $X^1 = (X_1,\ldots,X_n)'$, $X^2 = (Z_1,\ldots,Z_n)'$, and $Y=(Y_1,\ldots,Y_n)'$. The estimated relationship is to be a linear combination of X^0, X^1 and X^2. That is, $\hat{Y} = aX^0 + bX^1 + cX^2$.

1. (5%) What is the square of the distance between Y and $\hat{Y}$?

   (Hint: Note that in the problem set, you saw $\| Y - \hat{Y} \|^2 = \sum_{j-1}^{n} (Y_j - a - bX_j)^2$.)
2. (5%) Show that the distance minimizing parameter *a* must be $a = \overline{Y} - b\overline{X} - c\overline{Z}$ where $\overline{Y}, \overline{X}, \overline{Z}$ are the sample means. (You don't have to worry about second order conditions.)
3. (5%) Define $x = X - \overline{X}$, $y = Y - \overline{Y}$ and $z = Z - \overline{Z}$, and appeal to 2. to simply $\| Y - \hat{Y} \|^2$ so it does not contain *a*.

   (Hint: In the problem set, you saw $\| Y - \hat{Y} \|^2 = \sum_{j-1}^{n} (y_j - bx_j)^2$.)
4. (5%) Solve for the distance minimizing value of *b* and c.
5. (5%) Verify that this is a special case of the OLS formula $\beta = (W'W)^{-1}W'Y$

   where $W = [X^0, X^1, X^2]$ and $\beta = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$.