Student: Huong (Hanna) Nguyen

Program: MA in International Economics and Finance

Course: ECON 212-Survey of Advanced Econometrics Techniques

Instructor: Prof. Fournier – Spring 2014

Final Statistical and Econometric Paper

Dataset 1: A Multivariate Regression Model on Determinants of Writing Performance

Dataset 2: A Logit Regression Model on Factors Affecting the Usage of Computer

Dataset 3: A Hierarchical Linear Model on Factors Affecting Solder's Well-being

I. DATASET 1: WRITING SCORE

1.1. Objective of the Analysis

This paper analyzes some critical factors that determine writing score on the basis of data available on 220 observations. The analysis relies on a multiple linear regression model of writing score (variable write) as a function of gender, race, socio-economic status level, type of school, type of program, and scores on other academic studies. First, the paper would start with a brief overview of the dataset and descriptive statistics, then examine the relationship between dependent variable and the proposed explanatory variables as well as the relationship among independent variables. Finally, the paper would construct a multiple linear regression and run through some simulations to interpret how the factors affect writing scores.

1.2. Univariate Analysis:

The data set include 200 observations and 20 variables in total. Out of the sample observed, there were 54.5% female student and the remaining is male counterparts. A majority of the students (72.5%) reported themselves as white, the rest was non-white. Furthermore, almost half of the students (47.5%) were in the middle socio-economic group, 29% of the students were from the top bracket of the socio-economic status, and 23.5% of the students identified themselves as part of the lowest socio-economic status

Only about 16% of the students were educated in private schools while the remaining (84%) attended the public school system. The date shows that 52.5% of the students were in academic program and approximately 22.5% were in a general high school program while the remaining 25% were in an unconventional vocational program. These figures indicate that academic programs are still the most popular choice to most of the students.

Out of the 200 observed, the writing score had an average of 52.78, widely ranging from a minimum of 31 to a maximum of 67. The standard deviation of writing score is 9.47, meaning that on average these scores of writing test vary by roughly 9.47 from their mean score of 52.78. A median of 54 indicates a slight positive skew of the scores to the right, meaning that less than half of the sample score below and more than half score above the mean.

Additionally, it would be interesting to compare writing scores with other scores because students also took tests in the other subjects such as reading, math, science and social studies. The average scores were 52.23, 52.64, 51.85, and 52.40 in reading, math, science and social studies respectively, which shows fairly equal scores among those test yet lower than the mean of writing scores (54.5%). The social study had the largest standard deviation of 10.75, which was followed by the reading, science and math with the standard deviation of 10.25, 9.9, and 9.36 respectively.

1.3. Bivariate analysis

- **Correlation between writing scores and other subject scores**

Examining the correlation relationship between writing score and other subject score, we can see that there is a highly significant ($p<0.01$) and moderately positive correlation between the writing scores and scores in reading($r=0.59$), math($r=0.61$), science (0.57) and social studies($r=0.60$).

- **Relationship between Writing scores and Gender, Race**

In this sample, a female is likely to score 4.87 points higher on their writing than a male counterpart ($t=-3.65$, $df =169.70$, $P<=0.05$). It can also be seen that white people on average have higher writing scores than their non-white counterparts ($t=3.17$, $df =198$, $P<=0.05$). It is interesting to notice that both female and white student seem to have insignificant preference to enroll in the private school either ($P>0.05$).

Analyzing difference in writing scores between races, the ANOVA revealed a significant ($F=7.83$, $p<0.05$) difference between the races, but a closer analysis showed that the difference existed between Asian and Hispanic (Asians were on average 11 points higher), African-American and Asian (Asian were about 10 points higher), White and Hispanic (Whites scored 7.5 points higher) and Whites and African-American (Whites scored about 5 points higher).

- **Relationship between Writing Scores and Socio-economic Status**

The writing scores among the socio-economic classes are also significantly different from one another ($F=4.97$, $p<0.00$) and the appropriate tests revealed a difference of 5.29 points higher of high economic status students than low economic status students and roughly 3.98 points higher than middle economic students. There are not much of significant difference between middle and lower economic status students.

- **Relationship between Writing Score and Types of programs, and Type of School**

The writing scores among the three different types of programs were significantly different ($F=21.27$, $p<0.01$) as a one-way ANOVA revealed a difference of 4.92 points higher between academic and general and approximately 4.5 points higher between general and vocational programs. This is largely consistent with the expectation that students joining academic programs would have a higher writing score.

It was also expected and confirmed that students in private schools will score significantly ($t= -2.22$, $p<0.03$) higher on the writing test than those in private schools. The difference was about 3.28 points between the two groups.

1.4. Model – Multivariate Regress Analysis

To determine the relationship between the dependent variable writing scores and the independent variables gender, race, socio-economic status, type of school and educational program and comparable scores in other tests, a simple linear regression model is run as follows:

Writing= β_0 + β_1(female) + β_2(white01) + β_3(lowses) + β_4(midses) + β_5(privateschool01) + β_6(acad01) + β_7(general01) + β_8(read) + β_9(math) + β_{10}(science) + β_{11}(socst)

Given this model, it is expected that there is a positive relationship between all the dependent variable and the independent variables. We take several assumptions for this linear regression model to make simple interpretations. Some of the important assumptions are linearity, no measurement errors, homoscedastic errors (errors with constant variance) and lack of perfect multicollinearity. These assumptions are central to the classical linear regression model. We will test the presence of any violations of these assumptions and correct the model accordingly.

Running a multiple linear regression, we have the following result:

Table 1: Multivariate Regression Model for Writing Score

	Model 1		**Model 2**	
Variable	***Coefficient***	***Sig.***	***Coefficient***	***Sig.***
female	5.370066	***	5.428215	***
white01	0.0123389			
lowses	0.8537581			
midses	-0.1488739			
privateschool01	1.069896			
acad01	1.736465			
general01	0.6660995			
read	0.1140999			
math	0.2073124	*	0.2801611	***
science	0.2598395	***	0.2786543	***
socst	0.2164894	***	0.2681117	***
_cons	6.785551		6.568924	
** is p≤ 0.05 ,** is p≤ 0.01 , *** is p≤ 0.001*				
	R-squared	0.6102	***R-squared***	0.5940
	Adj R-squared	0.5874	***Adj R-squared***	0.5857
	AIC	1301.778	***AIC***	1295.905
	BIC	1341.358	***BIC***	1312.397

It is fairly surprising that the regression results of Model 1 are contradictory to our prior expectations. While we did expect the other subject scores to be important indicator of a writing

score, it was surprised to see that expected key variables such as race, Scio-economic status, type of school, academic program, and even reading score were not significant ($p<0.05$). A female was likely to score 5.37 points more than a male counterpart, holding other variables constant. Every unit increase in subjects such as math, science and social studies would individually increase the student's writing score by 0.20 points, 0.26 points and 0.21 points respectively, holding every else constant. This model only explained about 61% of the variations in writing scores and did not violate any of our OLS assumptions.

Model 1

Writing= 6.786 + 5.37(female) + 0.01(white01) + 0.85 (lowses) -0.14 (midses) + 1.06(privateschool01) + 1.73(acad01) + 0.66(general01) + 0.11(read) + 0.20(math) + 0.26(science) + 0.21(socst)

Dropping all insignificant variables and keeping only four significant ones, we run another multiple linear regression model expressed as follows:

Model 2

Writing= $\beta_{0'} + \beta_{1'}$(female) + $\beta_{9'}$(math) + $\beta_{10'}$(science) + $\beta_{11'}$(socst)

OR: Writing= 6.56 + 5.42(female) + 0.28(math) + 0.28(science) + 0.26(socst)

After dropping the non-significant variables, the second model fared better with decreasing AIC and BIC values ($AIC_{Model2}<AIC_{Model1}$, $BIC_{Model2}<BIC_{Model1}$). A female was likely to score 5.42 points higher than a male counterpart, holding every else constant. Each unit change in subjects such as math, science and social studies would individually increase the student's writing score by 0.28 points, 0.28 points and 0.26 points respectively, holding everything else constant. While the model did not show any multicollinearity, it did show slight heteroskedasticity ($chi^2=22.03$ $p=0.0549$) which we were able to correct using robust standard errors.

1.5. Conclusion

Examining the determinants of writing score in the sample, we run multiple linear regression models. At first, it is expected that all the explanatory variables including gender, race, socio-economic status level, type of school, type of program, and scores on other academic studies would have significant impacts on the writing scores. However, after dropping insignificant variables, the final model shows that only four factors, namely gender, math score, science, social studies, significantly determine writing score in this sample; these variables explain about 60% of the variation in writing scores. Though this is still a significant portion unexplained, this represents a good model, given the data available.

DATASET 2: COMPUTER USAGE

An Econometric Analysis on Factors Affecting the Usage of Computer

2.1. Purpose of the analysis

This paper aims to analyze some critical factors that determine the using of computer on the basis of data from the General Social Survey (GSS) in 2004. The analysis relies on a logit model developed to predict the probability of a person using a computer. The logit model with the dependent variable "usecomputer" is a function of a number of independent variables including college, female, marriage status, age, black, education and income. First, the paper would start with a brief overview of the dataset, and then examine the relationship between the dependent variable and the proposed explanatory variables as well as the interrelationship among independent variables. Finally, the paper would construct a logit model to generate statistical results for interpreting how the factors affect the probability of a person using a computer.

2.2. Data Overview

The data set includes more than 6,079 observations, but there are a number of missing values in independent variables. Surprisingly, nearly 36% of the respondents did not use a computer in 2004. There are more women than men in the dataset, since 56% of the population was female and about 44% was male. The age of the population was fairly old, with the mean of 59.46 and median of 57. Only 18% of participants reported themselves as African American and the majority of the participants (80%) was already married.

Regarding education level which is represented by the number of school year completed, the level of education ranges widely, from 0 years to 20 years. Most of the participants finished high school, two-year and four-year college program, which accounted for the highest percent of 32%, 11%, and 12% respectively. Only approximately 2% of the participants completed 20 years of educational training. The income range were also fairly different among respondents. Nearly 63% of them earned less than US$25,000 per year, and only 32% earned more than US$ 25,000 and about 5% of the participants refused to release information on their salary.

2.3. Interrelationship Among Variables

First examining the relationship between dependent variable (useacomputer) and the proposed explanatory variables, it is easy to notice that the higher level of education a person is, the more likely he/she has been used a computer. As expected, there was a highly significant relationship between computer proficiency and education level ($t = -35.63$, $df = 6064$, $p<0.001$). The income range was also found to have highly significant relationship with computer use ($t = -9.99$, $df = 4127$, $p<0.001$), indicating that the more a person can earn, the more likely he or she has used a computer before. Meanwhile, gender (being a female) was found to be statistically insignificant with the using of a computer (Pearson $chi2\ (1) = 1.3926$, $p>0.05$). In contrast, the black is less likely to use a computer, it has a negatively significant correlation with dependent variable "useacomputer" (Pearson $chi2\ (1) = 44.0463$; $p<0.001$). Likewise, the old people tended not to use a computer, as expected ($t = 30.1326$, $df = 6058$, $p<0.001$). Married people

were found to be less likely to have a computer before (Pearson chi2 (1) = 20.5532, p<0.001). This trend was understandable, since the people often get married when they are old or mature enough.

2.4. Model

To determine the relationship between the dependent variable, use a computer, and the independent variables including college, female, marriage status, age, black, education and income, a LOGIT regression is chosen. The reason for choosing a multivariate logistic model is that it serves well as a prediction model. Particularly, it gives the probability of an outcome occurring when there are only two possible outcomes. An initial model that includes all variables is run, then a final model dropping insignificant variables is developed as follows:

Table 2: Logit Model for Computer Usage

	Initial Model		**Final Model**	
Variable	**Coefficient**		**Coefficient**	
EDUC	0.389	***	0.390	***
RINCOME	0.082	***	0.081	
collegedummy01	-0.042			
female01	0.367	***	0.367	***
evermarried01	0.411	***	0.411	***
age	-0.040	***	-0.040	***
black01	-0.557	***	-0.557	***
_cons	-3.426	***	-3.485	
AIC	3832.173		3830.2	
BIC	3882.749		3874.454	
GOF HL (p)	13.72(0.09)		9.52(0.30)	

** is p≤ 0.05 ,** is p≤ 0.01 , *** is p≤ 0.001*

The final model illustrates that each additional increase in educational level would increase the log-odds of using a computer by 0.39 units, holding all other variables constant. For every unit increase in income, he or she is 0.081 more likely to have used a computer before, keeping other variables equally. Being a female and a married person actually is 0.367 and 0.411 respectively more likely to use a computer than the counterparts. In contrast, every unit increase in age, we expect a corresponding decrease of 0.04 units in the log-odds of using a computer, controlling all other variable unchanged. Finally, being a black actually will decrease the probability of using a computer by 0.557 units.

When comparing the AIC and BIC between the two models, we see that the final model have lower AIC and BIC. Additionally, the final model survive the goodness of fit test with higher adjusted R-square. Therefore, the final model is a well-grounded one, illustrating that **education, income level, gender, marital status, age and race are main determinants of whether a person has used a computer before.**

DATASET 3: WELL-BEING OF A SOLDIER

An Analysis on Factors Affecting Soldier's Well-being at Both Individual and Group Level

3.1. Introduction

This paper aims to examine some critical factors that determine the well-being of soldiers in the U.S. and European army at different clusters. We are given a data set of observations and some explanatory variables including the number of hours worked, cohesive scores, and leadership scale. The well-being of the soldier is the dependent and treated as continuous variable. First, the paper would start with a brief overview of the dataset, and then examine the relationship between the dependent variable and the proposed explanatory variables as well as the relationship among independent variables. Finally, the paper would construct hierarchical linear model to generate statistical results for interpreting how the factors affect the solders' well-being.

3.2. Data Overview

The dataset consists of 7,382 soldiers in army stationed in the U.S. and in Europe. First of all, the average well-being scores ranged from 0 to 5. The mean score among soldiers in the current sample was a 2.78, which was slightly above the midpoint of the scale of 2.5. Scores were fairly widely-spread (SD = 0.91) in a roughly normal distribution. There was some skew toward the bottom of the scale with more soldiers scoring low scores.

Regarding leadership, almost all soldiers rate leadership in the middle between 2 and 4 (mean = 2.89, SD = 0.77). They were also likely to rate company leadership a bit lower. The frequency that some soldiers ranked company leadership very low (less than 2) was about 2.5 times higher than the high ranks (greater than 4). In terms of cohesion, the average score ranged in the middle between 1 and 5, with a mean of 3.07 and SD = 0.87. The cohesion score skewed toward the high end of the scale (median = 3.13).

For the number of hours, the vast majority of soldiers reported working long hours, with a mean amount of daily hours of 11.30 hours and a narrow distribution (SD = 2.27). Specifically, nearly 50% of soldiers worked for at least 12 hours and more than two-thirds of soldiers reported that they worked more than the American norm of 8-hour work day. There are 34 soldiers reported that they even worked for 24 hours per days on a usual basis, which was so astonished and questionable. Only 2% reported an average workday of less than 8 hours. Taking such long working hours in consideration, the number of working hours would be a very critical factor affecting the well-being of a soldiers.

Additionally, company membership is another variable we need consider in examining its relationship with soldier's well-being. The sample include various different company sizes. There were 226 soldiers in the largest company but only 15 in the smallest, making the smallest company about $1/15^{th}$ the size of the largest company. Two-thirds of companies had at least 50 members while less than 1 out of 10 companies included 150 soldiers or more. Just 6 of 99 had fewer than 30 members.

3.3. Interrelationship Among Variables

As expected, the number of working hours and the well-being of soldiers have a negative but highly significant correlation ($r = -0.1632$, $p<0.001$). For every unit increase in the working hour would decrease the well-being of the soldier by 0.1632, holding everything constant. Interestingly, to the soldiers, the leadership scale are found to be highly significant with the sense of well-being ($r = 0.4257$, $p<0.001$), which means that an additional increase in leadership scale would increase soldier's well-being by 0.4257 unit, keeping all other variable equal. Likewise, the cohesion score also have a positive and significant linear correlation with the well-being status ($r = 0.2403$, $p<0.001$), indicating that soldiers would have their sense of well-being increased by 0.24 units for each increase in the cohesion score.

Furthermore, it is expected that the size of a soldier's company are related to his or her perception of cohesion and leadership. This is because soldier's well-being could also be associated with his or her company's size. Variation among other company-level factors (for example different combinations of individuals with different of personal characteristics, roles, and duties, different company cultures and duties) could also affect individual-level ratings of cohesion, leadership, and well-being.

3.4. Model

To determine soldiers' well-being at both individual and group level, six models were run to compare different results to determine the optimal model. The first model is just the normal regression model to estimate fix effects, then the next five models would focus on the random effects. The model 2 is a basic random effect model including a fixed part with estimation for each covariate. The random part which consists of variance in the error tem is divided into cluster (company) and individual (soldier) components.

For the model 3 and model 4, the company-level (or "cluster") means for each of the covariates were added to the fixed part of the model. Estimated individual-level effects of hours, cohesion, and leadership were therefore based on the difference between values for individual soldiers and the average values for their companies. Including cluster means allowed us to analyze difference between-company and within-company effects. This approach made it possible to model the contextual effects of being in different companies and address potential confounding at the company level. The fixed part of RE models 3 and 4 produced covariate point estimates that did not appear all that different from the OLS model. The difference between RE models 3 and 4 is that the cluster mean for cohesion was dropped in model 4.

Table 3: OLS Model and HLM for the Soldier's Well-being

	Model 1 - OLS	**Model 2 - RE**	**Model 3-RE**	**Model 4-RE**
		Random-intercept models		
	Regression	Random Effects	Random effects with cluster means	
		Fixed Part		
Variables				
constant	1.73 ***	1.53 ***	3.54 ***	3.50 ***

hours	-0.04	***	-0.03	***	-0.03	***	-0.03	***		
mn_hours					-0.12	***	-0.02	***		
cohesion	0.07	***	0.08	***	0.08	***	0.08	***		
mn-cohesion					-0.04					
leadership	0.45	***	0.47	***	0.47	***	0.47	***		
mn-leadership					-0.22	***	-0.24	***		
				Random Part						
leadership cohesion cluster-level var			0.02	***	0.009	***	0.009	***		
individual-level var			0.643	***	0.643	***	0.643	***		
AIC	17922.92		17808.56		17776.23		17774.41			
BIC	17950.55		17850.00		17838.39		17829.67			

** is p≤ 0.05 ,** is p≤ 0.01 , *** is p≤ 0.001*

Dropping the cluster mean for cohesion make the between and within effects for cohesion to be the same (β = 0.079, *SE* = 0.012). Thus, holding hours worked and leadership constant, each unit increase in cohesion score for soldiers in the same company was associated with an increase of 0.08 units in well-being score. Meanwhile, any additional unit increase in leadership score for soldiers in the same company was associated with an increase of 0.47 units in well-being score, keeping cohesion and hours worked equally. The contextual effect of leadership—the impact of a one-unit increase in a company's mean leadership score after controlling for individual soldier leadership scores—was a 0.243-point decrease in well-being. In short, increase in individual-level perception of leadership would enhance the sense of well-being while average company-level perception appeared to deteriorate the relationship when comparing soldiers between groups. For the cluster means for hours worked, it seemed to intensify soldier-level effects on well-being. Holding other covariates unchanged, a one-hour increase in work for individual soldiers in the same company was associated with a small decrease in well-being scores (β=-0.026, SE=0.004). After controlling for the hours worked by individual soldiers, the contextual effect of average hours worked within a company was larger—about a 0.12-point decrease in well-being.

Regarding model fit, RE model 4 featured better goodness-of-fit measures than RE model 3 (no cluster means) and the initial OLS model. Information criteria (AIC and BIC value) all decreased from OLS to RE model 1 and then further to RE model 4 (Table 2). The smaller values of these tests suggests there is a decrease in discrepancy between observed and expected values.

The final two models is built by adding random slopes for leadership and cohesion to the random intercepts model that included cluster means for the hours and leadership covariates (RE model 4). In the first random slope model (RE model 5), a single random slope, for leadership, was added. In the next model (RE model 6), a random slope for cohesion was also added but the estimated variance did not reach statistical significance (ψ = 0.002, *SE* = 0.002).

Table 4: HLM Model (Random Coefficient) for the Soldier's Well-being

	Model 5 -RE		**Model 6-RE**	
	Random-coefficient models			
Variables	**Random Coef.1**		**Random Coef. 2**	
	Fixed Part			
constant	3.31	***	3.29	***
hours	-0.03	***	-0.03	***
mn_hours	-0.11	***	-0.11	***
cohesion	0.08	***	0.08	***
mn-cohesion				
leadership	0.47	***	0.47	***
mn-leadership	-0.22	***	-0.22	***
	Random Part			
leadership	0.01	***	0.011	***
cohesion				
cluster-level var	0.121	***	0.114	***
individual-level var	0.638	***	0.636	***
AIC	17754.83		17759.15	
BIC	17823.90		17848.94	

** is p≤ 0.05 ,** is p≤ 0.01 , *** is p≤ 0.001*

In RE model 5, variance in the random slope for leadership was statistically significant (ψ = 0.010, *SE* = 0.004). Hence, for every one unit increase in leadership, in 95% of Army companies the projected impact of perceived leadership would range from about 0.47 to 0.6 additional units on the well-being scale. This represents a considerable amount of variation given the sample mean effect of leadership of about 0.47 from the fixed part of the model. The estimated correlation between random intercepts and slopes was negative and strongly linear (ρ = -0.975). As the mean well-being scores of an army company increase, therefore, it's expected that the impact of perceived leadership will decrease.

Based on the RE model 6, it can be stated that hours worked, perceived company leadership, and perceived company cohesion were all significantly associated with individual soldier well-being. On average, working longer hours tended to worsen well-being, keeping covariates unchanged. Increases in perceived company leadership and cohesion was likely to lead to enhanced well-being, holding hours worked equally. Perceived leadership was found to have a

positive and significant relationship with well-being. Based on the random part of the model, average score of well-being changed widely from one company to another. Moreover, the association between perceived leadership and well-being tended to vary across army companies. The estimated impact of perceived leadership on well-being could be up to 40% smaller or larger than the average for the overall sample, depending on which company that a soldier belonged to. The effect of perceived leadership was found to decline when the average well-being increased.

Comparing the two final models, the fixed part in model 5 has the estimates and standard errors fairly unchanged from those estimated in the RE model 4. Furthermore, the goodness-of-fit was significantly improved in the model 5; the values of AIC and BIC were much lower in model 5 than in model 6. **Therefore, the Model 5 would be the most well-grounded one for this army sample.**

3.5. Conclusion

The HLM show that the number of daily work hours, cohesion score and leadership capability are all affect the soldier's well-being significantly. However, it would be overlooked if we only examine those factors at individual level. Instead, we should investigate those factors together with company-level characteristics, since soldiers in this sample are associated themselves with their own army groups. The results in our models indicate that 3% of the variance in the model would be missed out if the difference among soldiers at various company-level were ignored.