1 (20 points) You are conducting an econometric investigation into the hourly wage rates of female and male employees in the San Francisco metropolitan area.
The sample data consist of observations for 13,118 employees on the following variables:

$ln(W_i)$ = the natural log of hourly wage rate of the i-th employee, in dollars per hour;

ED_i = years of formal education completed by the i-th employee, in years;

AGE_i = age of the i-th employee, in years;

$Female_i$ = dummy variable defined such that Female=1 if the i-th employee is female and Female=0 if the i-th employee is male;

The regression model you propose to use is.

Model1:

$$\ln(\mathrm{W})_i = \beta_1 + \beta_2 \mathrm{ED}_i + \beta_3 \mathrm{AGE}_i + \beta_4 \mathrm{AGE}_i^2 + \beta_5 \mathrm{ED}_i \mathrm{AGE}_i + \beta_6 \mathrm{Female}_i + u_i \quad \text{(Eq.1)}$$

However, your friend suggest you to extend the original model to allow the interaction term between the female variable and other variables by adding the "$\mathrm{Female}_i\mathrm{ED}_i$" "$\mathrm{Female}_i\mathrm{AGE}_i$" "$\mathrm{Female}_i\mathrm{AGE}_i^2$" "$\mathrm{Female}_i\mathrm{ED}_i\mathrm{AGE}_i$" to the equation. The new model is

Model2:

$$\begin{aligned}\ln(\mathrm{W})_i = {} & \beta_1 + \beta_2 \mathrm{ED}_i + \beta_3 \mathrm{AGE}_i + \beta_4 \mathrm{AGE}_i^2 + \beta_5 \mathrm{ED}_i \mathrm{AGE}_i + \beta_6 \mathrm{Female}_i \\ & + \beta_7 \mathrm{Female}_i \mathrm{ED}_i + \beta_8 \mathrm{Female}_i \mathrm{AGE}_i + \beta_9 \mathrm{Female}_i \mathrm{AGE}_i^2 + \beta_{10} \mathrm{Female}_i \mathrm{ED}_i \mathrm{AGE}_i + u_i\end{aligned} \quad \text{(Eq.2)}$$

where the β_j (j=1,2,3,...,10) are regression coefficients and u_i is a disturbance term.
Using the sample data describe above, your group computes OLS estimates of regression equation model1 and model2 . For each of the sample regression equations estimated on the sample data, the following table contains the OLS coefficient estimates (with estimated standard errors in parentheses below the coefficient estimates) and the summary statistics RSS (residual sum-of-squares), TSS (total sum-of- squares), and N (number of sample observations).

Table 1. OLS Sample Regression Equations for $ln(W)_i$ (standard errors in parentheses)

	Model 1	**Model 2**
Intercept (β_1)	−18.93 (3.013)	−23.07 (3.918)
$ED_i(\beta_2)$	0.3359 (0.1472)	0.0137 (0.1906)
$AGE_i(\beta_3)$	1.263 (0.1088)	1.436 (0.1436)
$AGE_i^2(\beta_4)$	−0.0155 (0.001029)	−0.0176 (0.001387)
$ED_i(\beta_5)$	0.02732 (0.003424)	0.03656 (0.004385)
$Female_i(\beta_6)$	−6.551 (0.1962)	1.416 (6.077)
$Female_iED_i(\beta_7)$	−	0.6049 (0.298)
$Female_iAGE_i(\beta_8)$	−	−0.2699 (0.2186)
$Female_iAGE_i^2(\beta_9)$		0.00303 (0.002054)
$Female_iED_iAGE_i(\beta_{10})$		−0.01935 (0.006962)
RSS=	1,638,981.08	1,613,827.86
TSS=	2,159,012.05	2,159,012.05
N=	13,118	13,118

1.1(4 points) Based on the regression result of **Model 1** on table 1, interpret carefully each of the slope coefficient estimate β_2.

1.2(4 points) Based on **Model 1**, holding other factors fixed, what is the approximate difference in the hourly wage rates of female and male employees in the San Francisco metropolitan area? Is this difference statistically significant at the 5 percent significance level?

1.3(4 points) Calculate the value of an appropriate goodness-of-fit measure (R^2 and adjusted-R^2) for each of the sample regression equations model1 and model 2 in the table. Which of the two sample regression models provides the best fit to the sample data? State the drawback of using R^2 to be a goodness-of-fit measure.

1.4(8 points) State the coefficient restrictions that regression of **Model 1** in the table imposes on regression of **Model 2**. Explain in words what the restrictions mean.

Use the estimation results given in the table to perform a test of these coefficient restrictions at the 5 percent significance level (i.e., for significance level $\alpha = 0.05$). State the null and alternative hypotheses, show how you calculate the required test statistic.

State the decision rule you use, and the inference you would draw from the test.

Based on the outcome of the test, which of the two regression models would you choose, Model (1) or Model (2)?

2 (15 points) Consider the a simple model of labor demand as follows:

You are conducting an empirical investigation the labor demand of Belgian firms . The sample data consists of 569 firms that includes information for 1996 on the following variables:

ln(labor) =natural log of total employment (number of worker);

ln(capital) = natural log of total fixed assets (in million euro);

ln(wage) =natural log of total wage costs divided by number of workers (in 1000 euro);

ln(output) = natural log of output (in million euro);

Your research assistant estimates a following regression model :

$$\ln(\text{labor})_i = \beta_1 + \beta_2 \ln(\text{wage})_i + \beta_3 \ln(\text{output})_i + \beta_4 \ln(\text{capital})_i \quad \text{(Eq.3)}$$

The estimation results for the model Eq.3 are given below.

Table 2.1 the regression result of model Eq.3

```
. reg ln_labor ln_capital ln_output ln_wage
```

Source	SS	df	MS	Number of obs =	569
				F(3, 565) =	(3)
Model	656.747035	(1)	218.915678	Prob > F =	0.0000
Residual	122.338812	565	.21652887	R-squared =	0.8430
				Adj R-squared =	0.8421
Total	(2)	568	1.37163001	Root MSE =	.46533

ln_labor	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ln_capital	-.0036975	.0187697	-0.20		-.0405644	.0331695
ln_output	.9900474	.0264103	(4)		.938173	1.041922
ln_wage	-.9277643	.0714046	-12.99		-1.068015	-.7875133
_cons	6.17729	.2462105	25.09		5.69369	6.660889

2.1 (5 points) From the table 2.1, fill in the information in boxes.

What is the d.f in box 1?

What is the sum of square in box 2?

What is the F-value in box 3?

What is the t-value in box 4?

Now, consider the following Stata command:

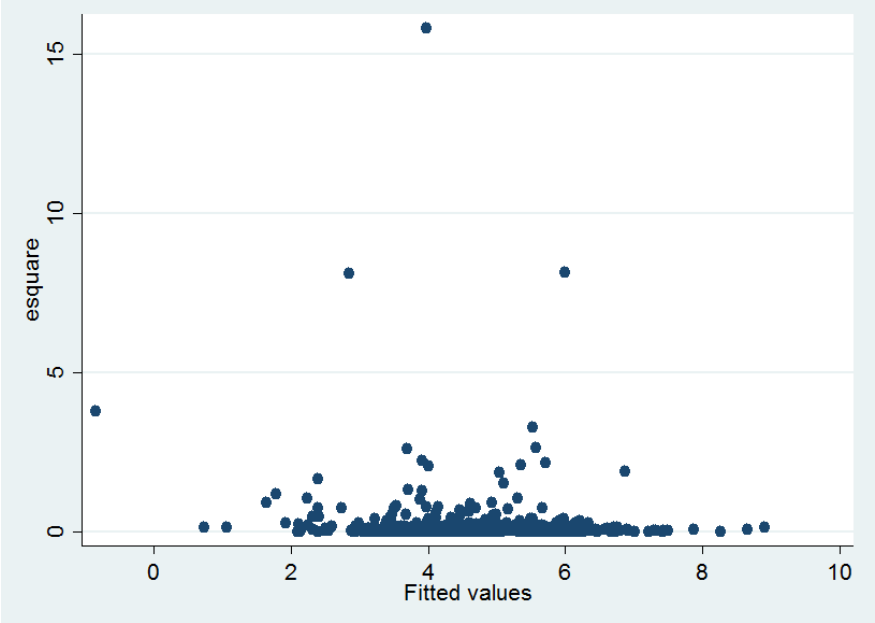
predict yhat if e(sample)

predict e if e(sample), resid

gen esquare = e^2

scatter esquare yhat

Figure 2.2 The relationship between u_i^2 and $\hat{Y}_i$ from the regression results of Eq.3



2.2 (2 points) From the figure 2.2 , is there any problem of Heteroskedasticity? Why or Why not?

Now, consider the following tests:

Test 1: Bruech-Pagan test

```
. predict yhat if e(sample)
(option xb assumed; fitted values)

. predict e if e(sample), resid

. gen esquare = e^2

. regress esquare ln_capital ln_output ln_wage
```

Source	SS	df	MS	Number of obs	=	569
				F(3, 565)	=	2.59
Model	6.12350437	3	2.04116812	Prob > F	=	0.0519
Residual	444.805265	565	.787265956	R-squared	=	0.0136
				Adj R-squared	=	0.0083
Total	450.92877	568	.793888679	Root MSE	=	.88728

esquare	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ln_capital	-.0629584	.0357899	-1.76	0.079	-.1332558	.0073391
ln_output	.0003251	.0503589	0.01	0.995	-.0985885	.0992386
ln_wage	.2453632	.1361537	1.80	0.072	-.0220659	.5127924
_cons	-.6159502	.4694717	-1.31	0.190	-1.538073	.3061728

2.3 (4 points) According to Bruech-Pagan test , does heteroskedasticity arise?
Conduct LM-test for checking heteroscedasticity at the 5 percent significance level (i.e., for significance level $\alpha = 0.05$)

Test 2: White-test

```
. estat imtest,white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(9)      =     58.54
         Prob > chi2  =    0.0000

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	58.54	9	0.0000
Skewness	6.62	3	0.0852
Kurtosis	2.69	1	0.1009
Total	67.85	13	0.0000

2.4 (4 points)
Based on Test2: White-test, does heteroskedasticity arise? **Conduct LM-test** for checking heteroscedasticity at the 5 percent significance level (i.e., for significance level $\alpha = 0.05$)

3. (15 points) This empirical illustration is based on the Puerto Rican employment rate, minimum wage, and other variables used by Castillo-Freeman and Freeman (1992) to study the effects of the U.S. minimum wage on employment in Puerto Rico.
A simplified version of their model is :

$$ln(prepop_t) = \beta_1 + \beta_2 ln(mincov_t) + \beta_3 ln(usgnp_t) + u_t, \quad \text{(Eq.4)}$$

where,
prepop: the employment rate in Puerto Rico during year t ;
usgnp: real U.S gross national product (in billions of dollars);
mincov: the importance of the U.S. minimum wage relative to average wages in Puerto Rico;

Using the data for the years 1950-1987, the estimation result is reported as below:

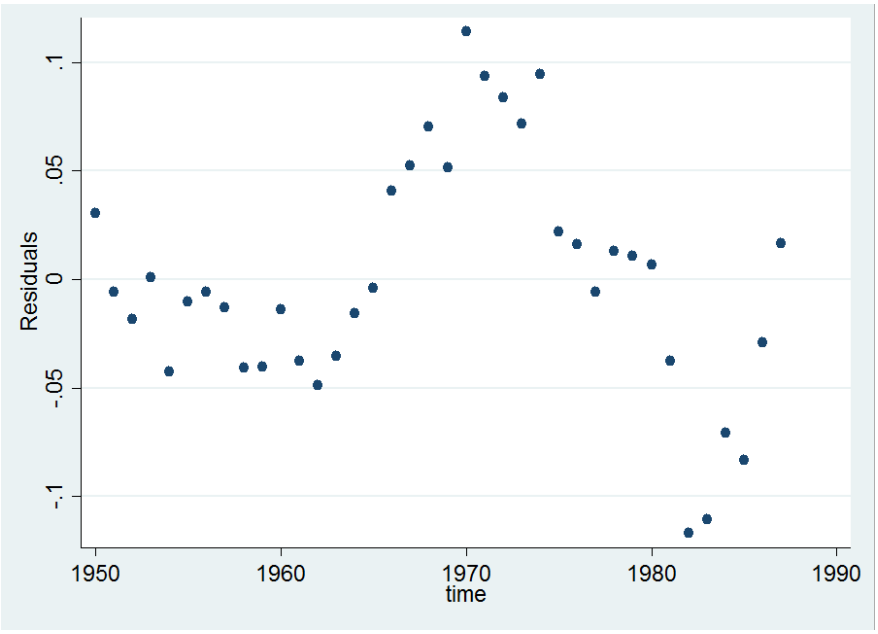
Table 3.1 the regression result of the effects of the U.S. minimum wage on employment in Puerto Rico

```
. regress lprepop lmincov lusgnp

      Source |       SS           df       MS      Number of obs   =        38
-------------+----------------------------------   F(2, 35)        =     34.04
       Model |  .211258366         2  .105629183   Prob > F        =    0.0000
    Residual |  .108600151        35  .003102861   R-squared       =    0.6605
-------------+----------------------------------   Adj R-squared   =    0.6411
       Total |  .319858518        37  .008644825   Root MSE        =    .0557

------------------------------------------------------------------------------
     lprepop |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lmincov |  -.1544442   .0649015    -2.38   0.023    -.2862011   -.0226872
      lusgnp |  -.0121888   .0885134    -0.14   0.891    -.1918806     .167503
       _cons |  -1.054423   .7654065    -1.38   0.177     -2.60828    .4994351
------------------------------------------------------------------------------
```

3.1 (2 points) Based on the result in Table 3.1, interpret carefully the slope coefficient estimate β_2.

Now, consider the following stata command:
predict uhat,resid
twoway (scatter uhat year) (line uhat year)

Figure 3.1



3.2 (2 points) From the Figure 3.1, is there the problem of autocorrelation? If yes, positive or negative autocorrelation ? Briefly explain how you detect it.

Test 3.1 Testing the AR(1) in disturbances of model Eq.4

```
. reg uhat L1.uhat,noconstant
```

Source	SS	df	MS		
Model	.074053296	1	.074053296	Number of obs	= 37
Residual	.033633102	36	.000934253	F(1, 36)	= 79.26
				Prob > F	= 0.0000
Total	.107686398	37	.002910443	R-squared	= 0.6877
				Adj R-squared	= 0.6790
				Root MSE	= .03057

uhat	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
uhat L1.	.8267816	.0928647	8.90	0.000	.6384432	1.01512

3.3 (4 points) Based on the Test 3.1, Is there **positive serial correlation** in the disturbances at the 5 percent level of significance? Show your work to receive full credits.

Test 3.2 Testing the AR(1) in disturbances of model Eq.4

```
. estat dwatson

Durbin-Watson d-statistic(  3,     38) =  .3396276

.
```

3.4 (4 points) Based on the Test 3.2, Is there **positive serial correlation** in the disturbances at the 5 percent level of significance? Show your work to receive full credits.

Table 3.3 New regression result of the effects of the U.S. minimum wage on employment in Puerto Rico

```
. prais lprepop lmincov lusgnp,rhotype(regress)

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.8268
Iteration 2:  rho = 0.8319
Iteration 3:  rho = 0.8323
Iteration 4:  rho = 0.8323
Iteration 5:  rho = 0.8323
Iteration 6:  rho = 0.8323

Prais-Winsten AR(1) regression -- iterated estimates
```

Source	SS	df	MS		
				Number of obs	= 38
				F(2, 35)	= 34.01
Model	.065096641	2	.03254832	Prob > F	= 0.0000
Residual	.033498278	35	.000957094	R-squared	= 0.6602
				Adj R-squared	= 0.6408
Total	.098594919	37	.002664728	Root MSE	= .03094

lprepop	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lmincov	-.1301007	.0486172	-2.68	0.011	-.2287989	-.0314025
lusgnp	-.0171942	.0831255	-0.21	0.837	-.1859479	.1515596
_cons	-.9773567	.6892484	-1.42	0.165	-2.376605	.4218919
rho	.8323472					

```
Durbin-Watson statistic (original)    0.339628
Durbin-Watson statistic (transformed) 1.763269
```

3.5 (3 points) Given Durbin-Watson result on Test 3.2, is it necessary to perform the regression in Table 3.3 instead of Table 3.1? Why?

4. (25 points) Give the complete answers to the following questions:

4.1 (5 points) What is the Variance Inflation factor (VIF)?

4.2 (5 points) How can we apply the VIF to measure of multicollinearity?

4.3 (15 points) State with reason whether the following statements are **true, false, or uncertain**

4.3.1 (5 points) Ceteris paribus, the higher the VIF is, the larger the variances of OLS estimators.

4.3.2 (5 points) You will not obtain a high R^2 value in a multiple regression if all the partial slope coefficients are individually statistically insignificant on the basis of the usual t test.

4.3.3 (5 points) If multicollinearity between two independent variables occurs, you should drop one of those independent variables from the model.

5. (25 points) Give the complete answers to the following questions:

5.1 (5 points) Based on OLS assumptions concerning properties of residual term, what problems can occur in case of cross-sectional data and time series data? and which assumptions are violated?

5.2 (5 points) What are the consequences of autocorrelation problems?

5.3 (10 points) How can we determine whether autocorrelation problem occurs and how to solve the problem?

5.4 (5 points) If the estimated results involve with multicollinearity and autocorrelation problems, which problem should be solved first? why?

The End of Exam