1 (18 points) You are conducting an empirical investigation into the median prices of houses in 506 communities of a large metropolitan area. The sample data consist of 506 observations on the following observable variables:

P_i = the median house price in community i, in dollars;

NOX_i = the level of nitrous oxide in the air of community i, in parts per 100 million;

$DIST_i$ = the weighted distance of community i from 5 employment centers, in miles;

$ROOMS_i$ = the average number of rooms per house in community i;

$STRAT_i$ = the average student-teacher ratio of schools in community i.

Your research assistant estimates the following model of median house prices on the sample of N = 506 observations. The OLS estimation results for the model are given below (with standard errors given in parentheses below coefficient estimates).

$$\ln(\mathrm{P})_i = \beta_1 + \beta_2 \ln(\mathrm{NOX})_i + \beta_3 \ln(\mathrm{DIST})_i + \beta_4 \mathrm{ROOMS}_i + \beta_5 \mathrm{START}_i + u_i \qquad \text{(Eq.1)}$$

OLS Estimates of Equation Eq.1:

$$\hat{\beta}_1 = 11.08 \quad \hat{\beta}_2 = -0.9535 \quad \hat{\beta}_3 = -0.1343 \quad \hat{\beta}_4 = 0.2545 \quad \hat{\beta}_5 = -0.05245$$
$$(0.3181) \quad (0.1167) \quad (0.04310) \quad (0.01853) \quad (0.005897)$$

$$\mathrm{RSS} = \sum_{i=1}^{N} \hat{u}_i^2 = 35.1835; \qquad \mathrm{TSS} = \sum_{i=1}^{N} \left(\ln P_i - \overline{\ln P}\right)^2 = 84.5822; \qquad N = 506$$

NOTE: standard errors in parentheses below coefficient estimates.
RSS is the Residual Sum-of-Squares and TSS is the total Sum-of-Squares from OLS estimation of regression equation Eq.1.

1.1(3 points) Interpret each of the slope coefficient estimates $\hat{\beta}_2$ and $\hat{\beta}_4$ in regression equationEq.1; that is, explain in words what the numerical values of the slope coefficient estimates $\hat{\beta}_2$ and $\hat{\beta}_4$ mean.

1.2(4 points) Use the estimation results for regression equation Eq.1 to test the individual significance of each of the slope coefficient estimates $\hat{\beta}_2$ for ln(NOX) and $\hat{\beta}_4$ for ROOMS. For each test, state the null and alternative hypotheses, show how you calculate the required test statistic, and draw the conclusion. Which of these two slope coefficient estimates are individually significant at the 5 percent significance level?

1.3(5 points) Use the estimation results for regression equation Eq.1 to test the joint significance of all the slope coefficient estimates at the 1 percent significance level (i.e., for significance level $\alpha = 0.01$). State the null and alternative hypotheses, show how you calculate the required test statistic, and state the decision rule you use, and the conclusion you would draw from the test.

1.4(6 points) Use the estimation results for regression equation Eq.1 to test the proposition that $\beta_2 = \beta_3$, i.e., the marginal effect of ln(NOX) on ln(P) equals the marginal effect of ln(DIST) on ln(P).
Explain in words what this proposition means. Perform the test at the 1 percent significance level.
State the null hypothesis H_0 and the alternative hypothesis H_1. Write the restricted regression equation implied by the null hypothesis H_0.
OLS estimation of this **restricted regression equation** yields a **Residual Sum-of-Squares value of RSS = 41.9532**. Use this information, together with the results from OLS estimation of equation Eq.1, to calculate the required test statistic, and state the decision rule you use, and the conclusion you would draw from the test.

2 (18 points) Consider the log monthly earnings equation as follows:

$$\begin{aligned}\log(\text{wage})_i \quad &= \quad \beta_1 + \beta_2\text{educ}_i + \beta_3\text{exper}_i + \beta_4\text{tenure}_i + \beta_5\text{married}_i \\ &\quad\quad \beta_6\text{black}_i + \beta_7\text{south}_i + \beta_8\text{urban}_i + u_i\end{aligned} \tag{Eq.2}$$

where educ, exper, and tenure are all relevant productivity characteristics. Married, black, south, and urban are qualitative variables.

$\log(\text{wage})_i$ =natural log of wage

educ_i =years of education

exper_i =years of work experience

tenure_i =years with current employer

Married_i =1 if married,

black_i = 1 if black,

south_i =1 if living in the south,

urban_i =1 if living in urban.

The estimation result model Eq.2 is reported as below:
Table 3.1 the regression result of model Eq.2

Source	SS	df	MS
Model	41.8377619	7	5.97682312
Residual	123.818521	927	.133569063
Total	165.656283	934	.177362188

Number of obs =	935
F(7, 927) =	44.75
Prob > F =	0.0000
R-squared =	0.2526
Adj R-squared =	0.2469
Root MSE =	.36547

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0654307	.0062504	10.47	0.000	.0531642	.0776973
exper	.014043	.0031852	4.41	0.000	.007792	.020294
tenure	.0117473	.002453	4.79	0.000	.0069333	.0165613
married	.1994171	.0390502	5.11	0.000	.1227801	.276054
black	-.1883499	.0376666	-5.00	0.000	-.2622717	-.1144281
south	-.0909036	.0262485	-3.46	0.001	-.142417	-.0393903
urban	.1839121	.0269583	6.82	0.000	.1310056	.2368185
_cons	5.395497	.113225	47.65	0.000	5.17329	5.617704

2.1 (2 points) Write out the regression equation for log monthly earning based on model Eq.2.

2.2 (4 points) Which of the coefficients are individually statistically significant at the 5 percent level of significance? State the critical value for hypothesis testing to receive full points.

2.3 (5 points) Holding other factors fixed, what is the approximate difference in monthly earnings between black and nonblacks? Is this difference statistically significant?

Next, we extend the original model to allow the interaction term between the married variable and the black variable by adding the "$\text{married}_i\text{black}_i$" to the equation. The new model is

$$\begin{aligned}\log(\text{wage})_i \quad = \quad & \gamma_1 + \gamma_2\text{educ}_i + \gamma_3\text{exper}_i + \gamma_4\text{tenure}_i + \gamma_5\text{married}_i \\ & \gamma_6\text{black}_i + \gamma_7\text{south}_i + \gamma_8\text{urban}_i + \gamma_9\text{married}_i\text{black}_i + u_i\end{aligned} \qquad \text{(Eq.3)}$$

The estimation result is reported as below:
Table 3.1 the regression result of model Eq.3

Source	SS	df	MS
Model	41.8849359	8	5.23561699
Residual	123.771347	926	.133662362
Total	165.656283	934	.177362188

Number of obs	=	935
F(8, 926)	=	39.17
Prob > F	=	0.0000
R-squared	=	0.2528
Adj R-squared	=	0.2464
Root MSE	=	.3656

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0654751	.006253	10.47	0.000	.0532034	.0777469
exper	.0141462	.003191	4.43	0.000	.0078837	.0204087
tenure	.0116628	.0024579	4.74	0.000	.006839	.0164866
married	.1889147	.0428777	4.41	0.000	.1047659	.2730635
black	-.24082	.0960229	-2.51	0.012	-.4292677	-.0523723
south	-.0919894	.0263212	-3.49	0.000	-.1436455	-.0403333
urban	.1843501	.0269778	6.83	0.000	.1314053	.2372948
marriedblack	.0613537	.1032747	0.59	0.553	-.1413259	.2640333
_cons	5.403793	.1141222	47.35	0.000	5.179825	5.627762

Note: marriedblack = married*black

2.4 (4 points) Write out the regression equation for log monthly earning based on model Eq.3. Does there appear to be a significant interaction effect among the new terms? Assess this with respect to both the black dummy variable and their interaction term and explained the results. What is the conditional expectation of $\log(\text{wage})_i$ for married blacks?

2.5 (3 points) What is the estimated wage differential between married blacks and married nonblacks?

3 (18 points) To assess the feasibility of a guaranteed annual wage (negative income tax), the Rand Corporation conducted a study to assess the response of labor supply (average hours of work) to increasing hourly wages.The data for this study were drawn from a national sample of 6,000 households with a male head earning less than $ 15,000 annually. The data were divided into 39 demographic groups for analysis. Because data for four demographic groups were missing for some variables, the data given in this example refer to only 35 demographic groups. Estimate the model Eq.4 reports in the Table 3.1

$$\begin{aligned}\text{Hours}_i &= \beta_1 + \beta_2\text{rates}_i + \beta_3\text{ERSP}_i + \beta_4\text{ERNO}_i + \beta_5\text{NEIN}_i \\ &\quad \beta_6\text{asset}_i + \beta_7\text{age}_i + \beta_8\text{DEP}_i + \beta_9\text{school}_i + u_i\end{aligned} \tag{Eq.4}$$

where

Hours_i = average hours worked during the year

rate_i = average hourly wage (dollars)

ERSP_i = average yearly earnings of spouse (dollars)

ERNO_i = average yearly earnings of other family members (dollars)

NEIN_i = average yearly nonearned income

assets_i = average family asset holdings (bank account, etc.) (dollars)

age_i = average age of respondent

DEP_i = average number of dependents

school_i = average highest grade of school completed

The estimation result model Eq.4 is reported as below:

Table 3.1 the regression result of model Eq.4

Source	SS	df	MS
Model	115385.114	8	14423.1393
Residual	24381.6287	26	937.754952
Total	139766.743	34	4110.78655

Number of obs =	35
F(8, 26) =	15.38
Prob > F =	0.0000
R-squared =	0.8256
Adj R-squared =	0.7719
Root MSE =	30.623

hrs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rate	-93.75259	47.14499	-1.99	0.057	-190.6605	3.155329
ersp	.0002254	.0382548	0.01	0.995	-.0784084	.0788592
erno	-.2149663	.0979392	-2.19	0.037	-.4162832	-.0136495
nein	.1572073	.5164059	0.30	0.763	-.9042802	1.218695
asset	.0155724	.0254048	0.61	0.545	-.0366479	.0677928
age	-.3486302	3.72233	-0.09	0.926	-7.999989	7.302729
dep	20.72802	16.88047	1.23	0.230	-13.97029	55.42633
school	37.32565	22.66518	1.65	0.112	-9.263304	83.9146
_cons	1904.577	251.9332	7.56	0.000	1386.721	2422.433

3.1 (2 points) Write out the regression equation for model Eq.4.

Table 3.2 the correlation matrix

	rate	ersp	erno	nein	asset	age	dep	school
rate	1.0000							
ersp	0.5717	1.0000						
erno	0.0590	-0.0410	1.0000					
nein	0.7018	0.2344	0.3591	1.0000				
asset	0.7789	0.2741	0.2922	0.9875	1.0000			
age	0.0442	-0.0153	0.7755	0.5024	0.4171	1.0000		
dep	-0.6014	-0.6929	0.0502	-0.5208	-0.5136	-0.0484	1.0000	
school	0.8813	0.5491	-0.2986	0.5392	0.6309	-0.3311	-0.6026	1.0000

3.2 (8 points) Is there evidence of multicollinearity in the data? How do you know? Explain your answers in detail and state the critical value for hypothesis testing to receive fully points. If there is the multicollinearity problem, what remedial action, if any, would you take?

Table 3.3 Variance Inflation Factors and Tolerance (VIF

Variable	VIF	1/VIF
asset	192.57	0.005193
nein	180.51	0.005540
school	25.40	0.039367
rate	17.08	0.058540
age	9.72	0.102869
dep	4.53	0.220992
ersp	3.50	0.285841
erno	3.14	0.318680
Mean VIF	54.55	

3.3 (3 points) Explain the outcome of Variance Inflation Factors and Tolerance (VIF).

3.4 State with reason whether the following statement are true or false.

a. (2.5 points) Despite perfect multicollinearity, OLS estimators are BLUE.

b. (2.5 points) In case of high multicollinearity, it is not possible to assess the individual significance of one or more partial regression coefficients.

4. (18 points) Consider Method of Generalized Least Square (GLS).

Through Ordinary Least Square (OLS), we minimize

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2.$$

Through Generalized Least Square (GLS), we minimize

$$\sum_{i=1}^{n} w_i \hat{u}_i^2 = \sum_{i=1}^{n} w_i (Y_i - \hat{\beta}_1^* - \hat{\beta}_2^* X_i)^2.$$

where $w_i = \frac{1}{\sigma_i^2}$

4.1 (5 points) Explain a rationale on the method of generalized least square (GLS). In other words; explain why GLS is **superior or more appropriate** than OLS in the presence of heteroscedasticity.

You are conducting an empirical investigation into the air quality in California for 2015. The sample data consists of 30 standard metropolitan statistical areas (SMSAs) on the following variables:

airq = indicator for air quality **(the lower the better)**;

vala = value added of companies (in 1000 US \$);

rain = amount of rain (in inches);

coas = dummy variable, 1 for SMSAs at the coast; 0 for others;

dens = population density (per square mile);

medi = average income per head (in US \$).

Your research assistant estimates a following regression model of air quality in California on the sample of N = 30 observations

$$\begin{aligned} \text{airq}_i &= \beta_1 + \beta_2 \text{vala}_i + \beta_3 \text{rain}_i + \beta_4 \text{coas}_i + \beta_5 \text{dens}_i \\ & \beta_6 \text{medi}_i + u_i \end{aligned} \tag{Eq.5}$$

The estimation results for the model Eq.5 are given below.

Table 4.1 the regression result of model Eq.5

```
. regress airq vala rain coas dens medi

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  5,    24) =    2.98
       Model |  8723.84625     5  1744.76925           Prob > F      =  0.0313
    Residual |  14058.4538    24  585.768906           R-squared     =  0.3829
-------------+------------------------------           Adj R-squared =  0.2544
       Total |     22782.3    29  785.596552           Root MSE      =  24.203

------------------------------------------------------------------------------
        airq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        vala |   .0008834   .0022562     0.39   0.699    -.0037731    .0055399
        rain |   .2506988   .3435183     0.73   0.473     -.458288    .9596857
        coas |   -33.3983   10.45752    -3.19   0.004    -54.98156   -11.81504
        dens |  -.0010734   .0016233    -0.66   0.515    -.0044237    .0022769
        medi |   .0005545   .0008503     0.65   0.521    -.0012003    .0023093
       _cons |   111.9347   15.33179     7.30   0.000     80.29141    143.5779
------------------------------------------------------------------------------

.
```

4.2 (3 points) From the table 4.1, interpret each of the slope coefficient estimates $\hat{\beta}_2$ and $\hat{\beta}_4$ in regression model Eq.5.

Now, consider the following Stata command:

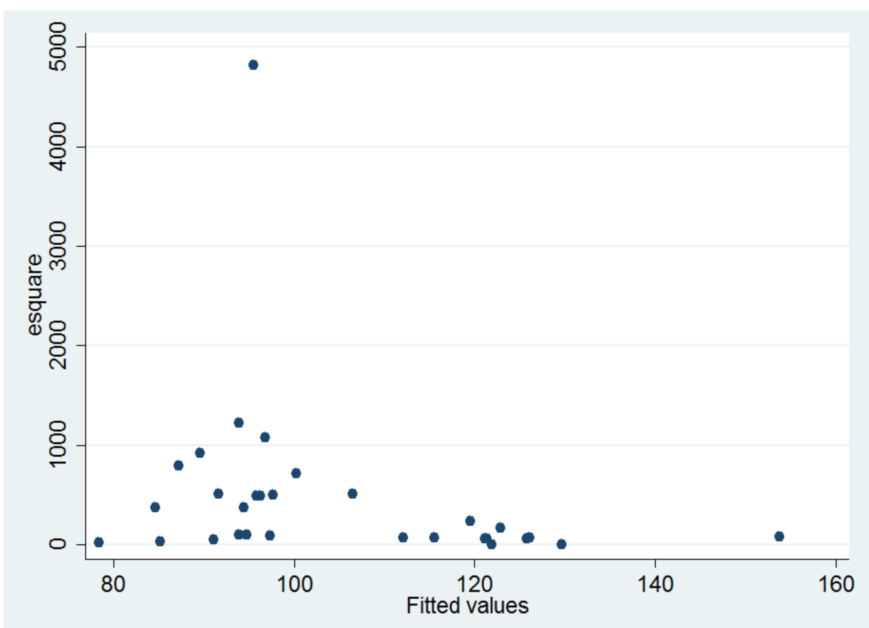
predict yhat if e(sample)

predict e if e(sample), resid

gen esquare $= e^2$

scatter esquare yhat

Figure 4.1 The relationship between u_i^2 and $\hat{Y}_i$ from the regression results of Eq.5



4.3 (2 points) From the figure 4.1 , is there any problem of Heteroskedasticity? Why or Why not?

Now, consider the following tests:

Test 1: Bruech-Pagan test

```
. predict e,resid

. predict yhat
(option xb assumed; fitted values)

. gen esquare =e^2

. scatter esquare yhat

. regress esquare vala rain coas dens medi
```

Source	SS	df	MS
Model	2406263.41	5	481252.682
Residual	20571541	24	857147.543
Total	22977804.4	29	792338.084

Number of obs	=	30
F(5, 24)	=	0.56
Prob > F	=	0.7284
R-squared	=	0.1047
Adj R-squared	=	-0.0818
Root MSE	=	925.82

esquare	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vala	.0001125	.0863058	0.00	0.999	-.1780138	.1782389
rain	-1.001573	13.14058	-0.08	0.940	-28.12239	26.11925
coas	570.1359	400.0307	1.43	0.167	-255.4869	1395.759
dens	-.0406206	.0620955	-0.65	0.519	-.1687795	.0875383
medi	-.004736	.0325247	-0.15	0.885	-.0718637	.0623917
_cons	220.2809	586.4857	0.38	0.711	-990.166	1430.728

4.4 (4 points) According to Bruech-Pagan test , does heteroskedasticity arise? Conduct F-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(19)      =      8.61
         Prob > chi2   =    0.9795

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

Source	chi2	df	p
Heteroskedasticity	8.61	19	0.9795
Skewness	2.32	5	0.8028
Kurtosis	0.81	1	0.3673
Total	11.74	25	0.9885

4.4 (4 points) According to White-test , does heteroskedasticity arise? Conduct LM-test for checking heteroscedasticity at the 5 percent significance level (i.e., for significance level $\alpha = 0.05$)

5. (18 points) This empirical illustration is based on one of the funding articles on autocorrelation, viz. Hildreth and Lu (1960). The data used in this study are time series data with 30 four-weekly observations from 18 March 1951 to 11 July 1953 on the following variables :

consump: consumption of ice cream per head (in pints);
income: average family income per week (in US Dollars);
price: price of ice cream (per pint);
temp: average temperature (in Fahrenheit).

The model used to explain consumption of ice cream is a linear regression model which is represented as:

$$ln(consump_t) = \beta_1 + \beta_2 income_t + \beta_3 ln(price_t) + \beta_4 temp_t + u_i \qquad \text{(Eq.6)}$$

The estimation result is reported as below:

Table 5.1 the regression result of the demand for ice cream

```
. tsset time
        time variable:  time, 1 to 30
                delta:  1 unit

. gen ln_consump =ln(cons)

. gen ln_price=ln(price)

. regress ln_consump income ln_price temp
```

Source	SS	df	MS
Model	.678880551	3	.226293517
Residual	.237165841	26	.009121763
Total	.916046392	29	.031587807

Number of obs =	30
F(3, 26) =	24.81
Prob > F =	0.0000
R-squared =	0.7411
Adj R-squared =	0.7112
Root MSE =	.09551

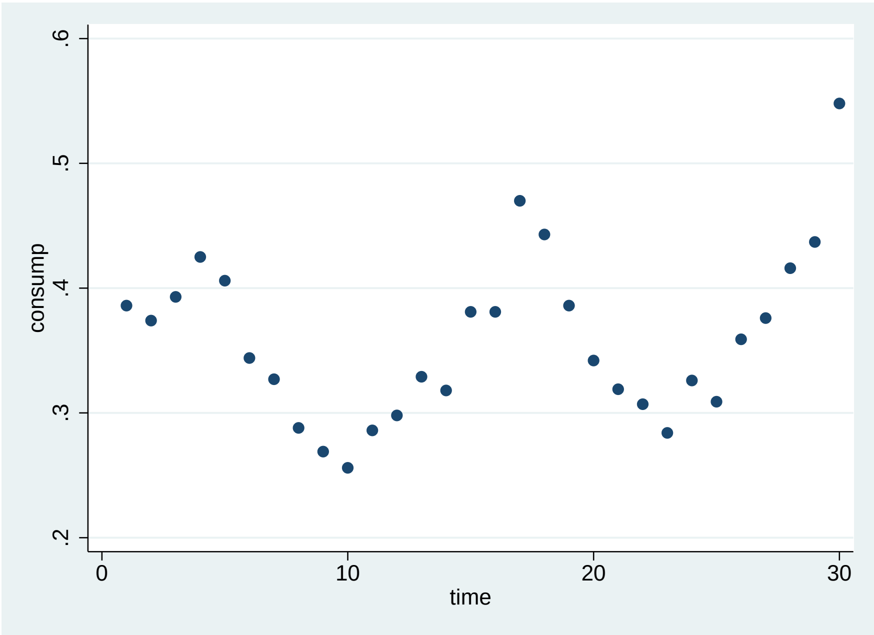
ln_consump	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.0093556	.0030391	3.08	0.005	.0031086	.0156026
ln_price	-.5954412	.5953848	-1.00	0.326	-1.819272	.6283898
temp	.0095892	.0011554	8.30	0.000	.0072142	.0119643
_cons	-3.069371	.7714954	-3.98	0.000	-4.655203	-1.48354

5.1 (4 points) Based on the regression result of the demand for ice cream on table 5.1, interpret carefully each of the slope coefficient estimates β_2 and β_3.

Now, consider the following stata command:

scatter consump time

Figure 5.1



5.2 (4 points) From the Figure 5.1, is there the problem of autocorrelation? If yes, positive or negative autocorrelation ? Briefly explain how you detect it.

Test 5.1 Durbin-Watson test on model Eq.6

```
. estat dwatson

Durbin-Watson d-statistic(  4,    30) =  .9826207

.
```

5.3 (5 points) Based on the Test 5.1, Is there positive serial correlation in the disturbances at the 5 percent level of significance? Show your work to receive full credits.

Table 5.2

```
. newey ln_consump income ln_price temp,lag(1)

Regression with Newey-West standard errors          Number of obs  =        30
maximum lag: 1                                      F(  3,     26) =     25.05
                                                    Prob > F       =    0.0000
```

ln_consump	Coef.	Newey-West Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.0093556	.0033274	2.81	0.009	.0025161	.0161951
ln_price	-.5954412	.6017979	-0.99	0.332	-1.832454	.6415722
temp	.0095892	.0011416	8.40	0.000	.0072427	.0119358
_cons	-3.069371	.6683789	-4.59	0.000	-4.443244	-1.695499

5.4 (5 points) Given Durbin-Watson result on Test 5.1, is it necessary to perform the regression in Table 5.2 instead of Table 5.1? Why? What is the limitation of using the technique in table 5.2?

The End of Exam