1. (10 points) You are conducting an empirical investigation into the prices of houses in a single large metropolitan area. The sample data consist of 88 observations on the following observable variables:

P_i = the selling price of house i, in thousands of dollars;

HS_i = the house size of house i, in hundreds of square feet;

YS_i = the yard size of house i, in hundreds of square feet;

DC_i = a dummy variable defined such that $DC_i = 1$ if house i is a colonial-style house, and $DC_i = 0$ if house i is not a colonial-style house.

Your research assistant estimates two alternative regression models of house prices on the sample of N = 88 observations for recently sold houses in a single large metropolitan area.

The estimation results for the two models are given below.

Model 1

$$P_i = \beta_1 + \beta_2 HS_i + \beta_3 YS_i + \beta_4 DC_i + U_i$$

OLS Estimates of Model 1

$$\begin{array}{rcl} \widehat{P}_i & = & -5.295 + 13.236\ HS_i + 0.211\ YS_i + 19.123\ DC_i \\ se & = & (24.772)\quad (1.136)\qquad\quad (0.064)\qquad\quad (13.898) \end{array} \qquad \text{(Eq.1)}$$

$$RSS = 302,371 \quad TSS = 917,855 \quad N = 88$$

Model 2

$$\ln(P_i) = \alpha_1 + \alpha_2 \ln(HS_i) + \alpha_3 \ln(YS_i) + \alpha_4 DC_i$$

$$\begin{aligned} \widehat{\ln(P_i)} &= 2.639 + 0.750\ \ln(HS_i) + 0.168\ \ln(YS_i) + 0.066\ DC_i \\ se &= (0.244)\quad (0.081)\qquad\quad (0.038)\qquad (0.043) \end{aligned}$$

(Eq.2)

$$RSS = 2.843 \quad TSS = 8.018 \quad N = 88$$

1.1 (2.5 points) Interpret each of the slope coefficient estimates β_2 and β_4 in regression model 1 (Eq.1).

1.2 (2.5 points) Interpret each of the slope coefficient estimates α_2 and α_4 in regression model 2 (Eq.2).

1.3 (5 points) Use the estimation results for regression model 1 to test the individual significance of each of the slope coefficient estimates β_2 for HS_i and β_4 for DC_i . For each test, state the null and alternative hypotheses, and show how you calculate the required test statistic. Which of these two slope coefficient estimates are individually significant at the 5 percent significance level?

1.4 (5 points) Use the estimation results for regression model 1 to test the joint significance of all the slope coefficient estimates at the 1 percent significance level (i.e., for significance level $\alpha = 0.01$).

1.5 (5 points) Use the estimation results for regression equation model 2 to test the proposition that $\alpha_2 > \alpha_3$, i.e., to test the proposition that the marginal effect of $\ln(HS_i)$ on $\ln P_i$ is greater than the marginal effect of $\ln(YS_i)$ on $\ln P_i$ at the 5 percent significance level. [Note: $\text{cov}(\alpha_2,\alpha_3\ )$ = -0.001]

2. (15 points) Consider the following model for the % students satisfactory of 4^{th} grade math from 1692 schools in Thailand :

$$\text{math4} = \beta_1 + \beta_2 \text{lunch} + \beta_3 \ln(\text{enroll}) + \beta_4 \ln(\text{exppp}) + u \quad \text{(Eq.3)}$$

where
math4 = % students satisfactory of 4^{th} grade math

lunch = % students eligible for free or reduced lunch

enroll = school enrollment

exppp = expenditures per pupil

Table 2.1 reported the regression results of Eq.3

```
. regress math4 lunch lenroll lexppp
```

Source	SS	df	MS	Number of obs	=	1,692
				F(3, 1688)	=	334.57
Model	235030.659	3	78343.5531	Prob > F	=	0.0000
Residual	395268.516	1,688	234.163813	R-squared	=	0.3729
				Adj R-squared	=	0.3718
Total	630299.175	1,691	372.737537	Root MSE	=	15.302

math4	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lunch	-.4487434	.014642	-30.65	0.000	-.4774618	-.420025
lenroll	-5.399152	.9404116	-5.74	0.000	-7.243648	-3.554657
lexppp	3.524744	2.097846	1.68	0.093	-.5899086	7.639397
_cons	91.93246	19.9617	4.61	0.000	52.78018	131.0847

Note:
lenroll=ln(enroll)
lexppp =ln(exppp)

2.1 (3 points) Interpret carefully each of the slope coefficient estimates β_2 and β_3 in Table 2.1

Now, consider the following Stata command:

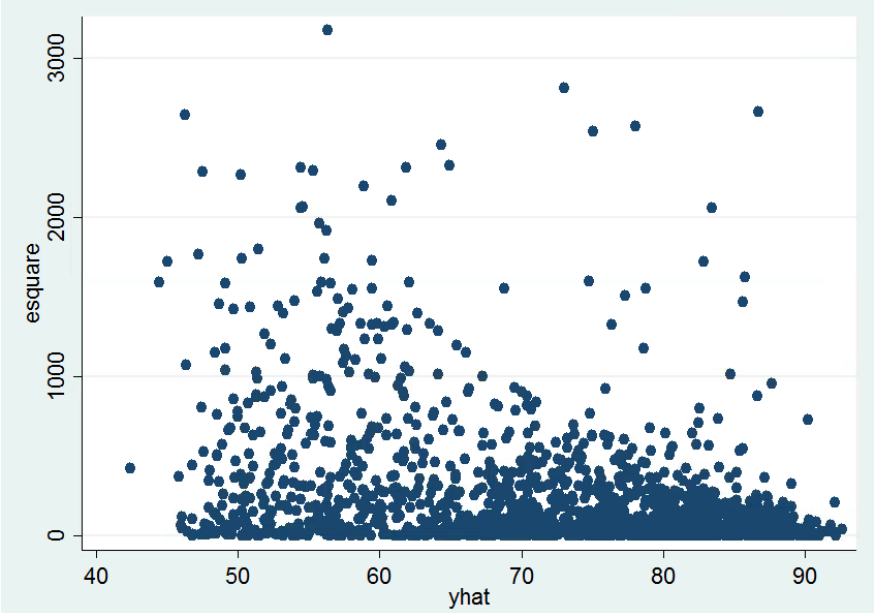
predict yhat if e(sample)

predict e if e(sample), resid

gen esquare = e^2

scatter esquare yhat

Figure 2.1 The relationship between u_i^2 and $\hat{Y}_i$ from the regression results of Eq.3



2.2 (4 points) From the figure 2.1, is there the problem of Heteroskedasticity? Why or Why not?

Now, consider the following tests:

Test 1: Bruech-Pagan test

```
estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
      Ho: Constant variance
      Variables: fitted values of math4

      chi2 (1)      = 320.38
      Prob > chi2   = 0.0000
```

Test 2: White's test

```
estat imtest, white

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

      chi2(9)       =   264.54
      Prob > chi2   =   0.0000
```

2.3 (8 points) According to Bruech-Pagan test and White's test, does heteroskedasticity arise? Fully explain with statistical supports at the 5 percent significance level (i.e., for significance level $\alpha = 0.05$)

3 (15 points) consider the following model:

$$\text{bwghtlbs} = \beta_1 + \beta_2 \text{cigs} + u \qquad \text{(Eq.4)}$$

where the dependent variable, infant birth weight in pounds (bwghtlbs), and an explanatory variable, average number of cigarettes the mother smoked per day during pregnancy (cigs). The following simple regression was estimated using data on n=1,388 births:

Table 3.1

```
. regress bwghtlbs cigs

      Source |       SS           df       MS      Number of obs   =     1,388
-------------+----------------------------------   F(1, 1386)      =     32.24
       Model |  51.0172632         1  51.0172632   Prob > F        =    0.0000
    Residual |  2193.55977     1,386  1.58265495   R-squared       =    0.0227
-------------+----------------------------------   Adj R-squared   =    0.0220
       Total |  2244.57703     1,387  1.61829634   Root MSE        =     1.258

------------------------------------------------------------------------------
    bwghtlbs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.0321108   .0056557    -5.68   0.000    -.0432054   -.0210161
       _cons |   7.485744   .0357713   209.27   0.000     7.415572    7.555915
------------------------------------------------------------------------------
```

Since higher income generally results in access to better prenatal care, as well as better nutrition for the mother, we therefore added one more explanatory variable, annual family income (faminc) to the Eq.4. The new result is reported as below:

Table 3.2

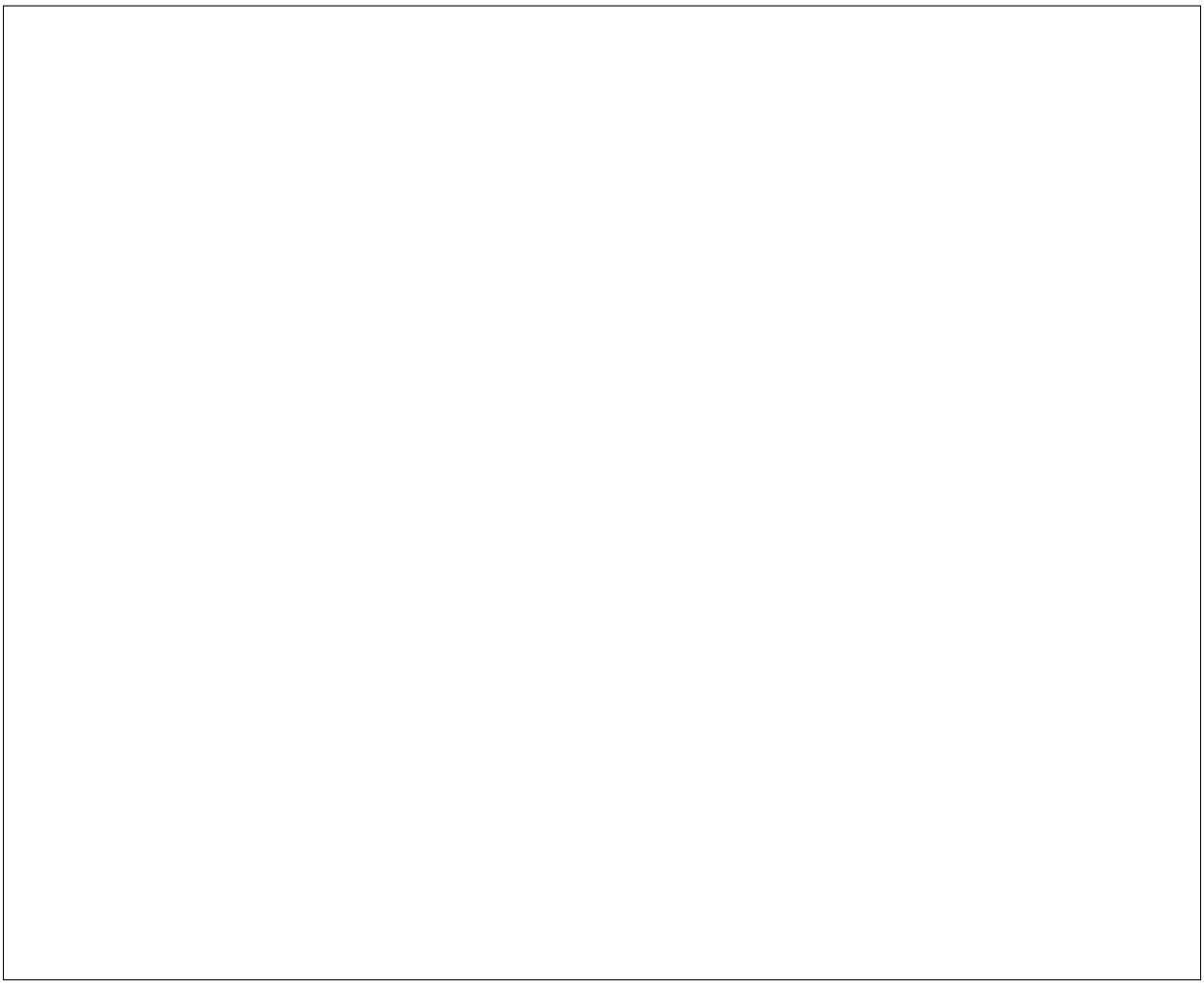
$$\text{bwghtlbs} = \beta_1 + \beta_2 \text{cigs} + \beta_3 \text{faminc} + u \qquad \text{(Eq.5)}$$

```
. regress bwghtlbs cigs faminc
```

Source	SS	df	MS	Number of obs	=	1,388
				F(2, 1385)	=	21.27
Model	66.8992533	2	33.4496266	Prob > F	=	0.0000
Residual	2177.67778	1,385	1.57233052	R-squared	=	0.0298
				Adj R-squared	=	0.0284
Total	2244.57703	1,387	1.61829634	Root MSE	=	1.2539

bwghtlbs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cigs	-.028963	.0057236	-5.06	0.000	-.0401907	-.0177352
faminc	.0057978	.0018242	3.18	0.002	.0022192	.0093764
_cons	7.310883	.0655615	111.51	0.000	7.182273	7.439494

3.1 (5 points) How would you choose between the Eq. 4 and Eq. 5 ? Which statistical test would you use to answer this question ? Show the necessary calculations.



Next, we extend the model in Eq.5 to allow infant birth weight in pounds (bwghtlbs) to depend on the average price of cigarettes (cigprice) and the cigarette taxes (cigtax). Therefore, we obtain the following result:

Table 3.3

```
. regress bwghtlbs cigs faminc cigtax cigprice
```

Source	SS	df	MS	Number of obs	=	1,388
				F(4, 1383)	=	11.56
Model	72.6029938	4	18.1507484	Prob > F	=	0.0000
Residual	2171.97404	1,383	1.57048014	R-squared	=	0.0323
				Adj R-squared	=	0.0295
Total	2244.57703	1,387	1.61829634	Root MSE	=	1.2532

bwghtlbs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cigs	-.029325	.0057234	-5.12	0.000	-.0405524	-.0180976
faminc	.0057549	.0018481	3.11	0.002	.0021295	.0093803
cigtax	.0092232	.0090334	1.02	0.307	-.0084975	.0269439
cigprice	-.0008785	.0069032	-0.13	0.899	-.0144204	.0126633
_cons	7.247245	.7450263	9.73	0.000	5.785741	8.708749

Table 3.4
corr bwghtlbs cigs faminc cigtax cigprice

	bwghtlbs	cigs	faminc	cigtax	cigprice
bwghtlbs	1.0000				
cigs	-0.1508	1.0000			
faminc	0.1089	-0.1730	1.0000		
cigtax	0.0478	0.0294	0.0179	1.0000	
cigprice	0.0492	0.0097	0.0955	0.8759	1.0000

3.2 (7 points) Is there multicollinearity in regression result Table 3.3? How do you know? Provide full explanation to award the full point.

3.3 (3 points) How do you describe OLS estimators when multicollinearity exists? Do they satisfy "BLUE" properties? Specify a remedial measure.

4. (20 points) Consider the Static Phillips curve model that explained **the inflation-unemployment trade-off** in the United States given by:

$$\text{inf}_t = \beta_1 + \beta_2 \text{unem}_t + u_t \quad \text{(Eq.6)}$$

where:
inf is the annual inflation rate.

unem is the unemployment rate.

The estimation result using the data from 1948 to 2003 (with 56 observations) is reported as below:

Table 4.1 the regression results of Eq.5

```
. regress inf unem
```

Source	SS	df	MS	Number of obs	=	56
				F(1, 54)	=	3.58
Model	31.599858	1	31.599858	Prob > F	=	0.0639
Residual	476.815691	54	8.8299202	R-squared	=	0.0622
				Adj R-squared	=	0.0448
Total	508.415549	55	9.24391907	Root MSE	=	2.9715

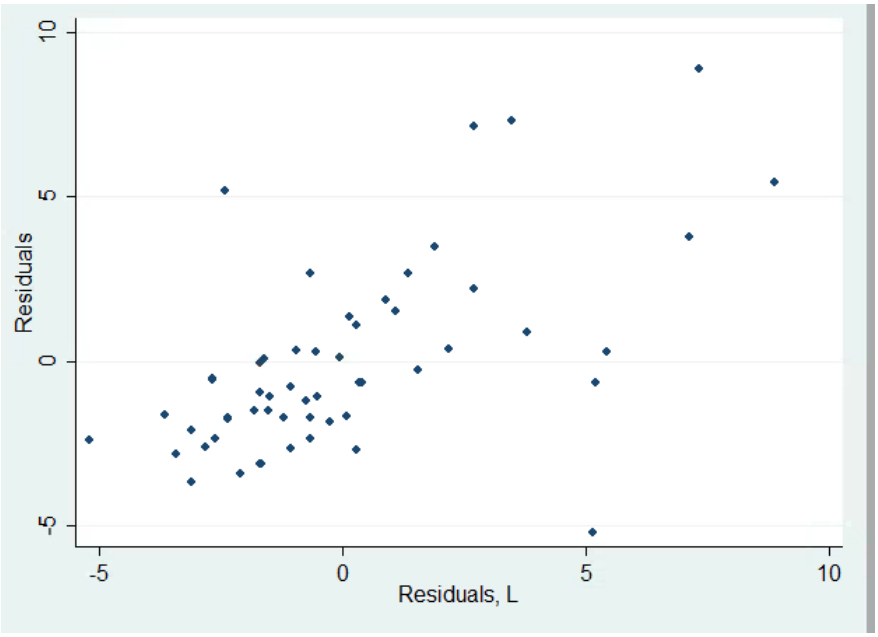
inf	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
unem	.5023782	.2655624	1.89	0.064	-.0300424	1.034799
_cons	1.053566	1.547957	0.68	0.499	-2.049901	4.157033

4.1 (5 points) Based on the Phillips curve theory, what should be the expected sign of β_2 of Eq.5? Compared with the estimation result in the table 4.1, do we get the result as expected from the Phillip curve theory? Why? Test the individual significance of the slope coefficient estimates β_2 at the 5 percent significance level (i.e. $\alpha = 0.05$)

Now, consider the following stata command:

predict e if e(sample), resid
gen uhat = e
gen esquare = e^2
scatter uhat L1.uhat

Figure 4.1



4.2 (5 points) From the figure 4.1, is there the problem of Autocorrelation? Briefly explain how you detect it.

Test 4.1

```
. tsset year
        time variable:  year, 1948 to 2003
                delta:  1 unit

. estat dwatson

Durbin-Watson d-statistic(  2,    56) =  .8014823
```

4.3 (5 points) Based on the test 4.1, Is there positive serial correlation in the disturbances at the 5 percent level of significance? Show your work to receive full credits.

Table 4.2

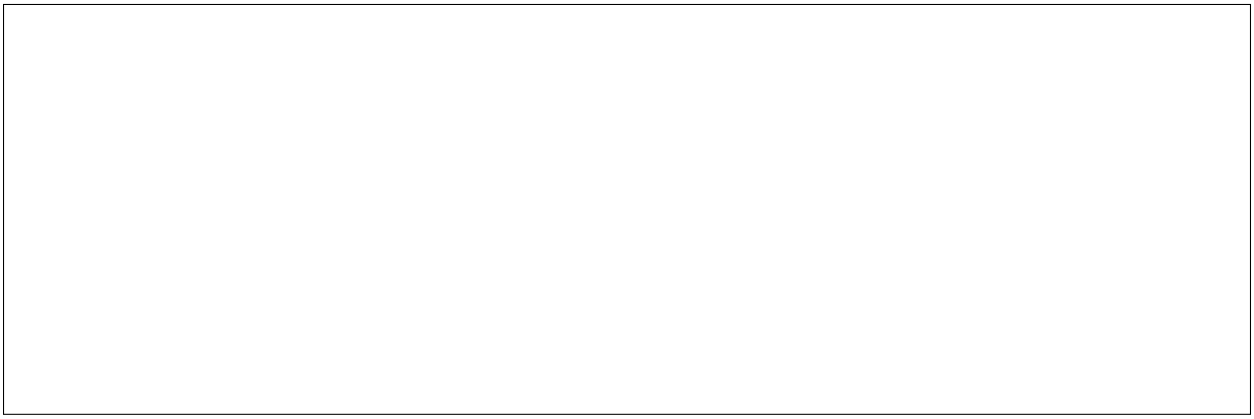
```
. prais inf unem , rhotype(regress)

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.5721
Iteration 2:  rho = 0.7350
Iteration 3:  rho = 0.7792
Iteration 4:  rho = 0.7871
Iteration 5:  rho = 0.7883
Iteration 6:  rho = 0.7885
Iteration 7:  rho = 0.7885
Iteration 8:  rho = 0.7885
Iteration 9:  rho = 0.7885

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

Source	SS	df	MS	Number of obs	=	56
				F(1, 54)	=	8.39
Model	38.377534	1	38.377534	Prob > F	=	0.0054
Residual	246.917431	54	4.57254502	R-squared	=	0.1345
				Adj R-squared	=	0.1185
Total	285.294965	55	5.18718118	Root MSE	=	2.1384

inf	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
unem	-.7139659	.2897858	-2.46	0.017	-1.294951	-.1329804
_cons	7.999443	2.048343	3.91	0.000	3.892762	12.10612
rho	.7885234					

```
Durbin-Watson statistic (original)    0.801482
Durbin-Watson statistic (transformed) 1.913928
```

4.4 (5 points) Given Durbin-Watson result on question 4.2, is it necessary to perform the regression in Table 4.2 instead of Table 4.1? Why?



5 (20 points) Consider a wage determination model as follows:

$$\begin{aligned}\log(\text{wage}) &= \beta_1 + \beta_1 \text{edu} + \beta_2 \text{exper} + \beta_3 \text{tenure} + \beta_4 \text{married} \\ &\quad \beta_5 \text{black} + \beta_6 \text{south} + \beta_7 \text{urban} + u\end{aligned} \tag{Eq.7}$$

where educ, exper, and tenure are all relevant productivity characteristics. Married, black, south, and urban are qualitative variables.
Married =1 if married,
black = 1 if black,
south=1 if living in the south,
and urban =1 if living in urban.

The estimation result is reported as below:

```
. gen logwage=log(wage)

. reg logwage edu exper tenure married black south urban
```

Source	SS	df	MS	Number of obs =	935
				F(7, 927) =	44.75
Model	41.8377619	7	5.97682312	Prob > F =	0.0000
Residual	123.818521	927	.133569063	R-squared =	0.2526
				Adj R-squared =	0.2469
Total	165.656283	934	.177362188	Root MSE =	.36547

logwage	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

5.1 (8 points) Holding other factors fixed, what is the approximate difference in monthly salary between blacks and nonblacks? Is this difference statistically significant at the 5 percent level of significance? Show your work to award the full point.

Next, we extend the original model to allow the return to education to depend on race by adding the interaction "blackeduc" to the equation and obtain the following result:

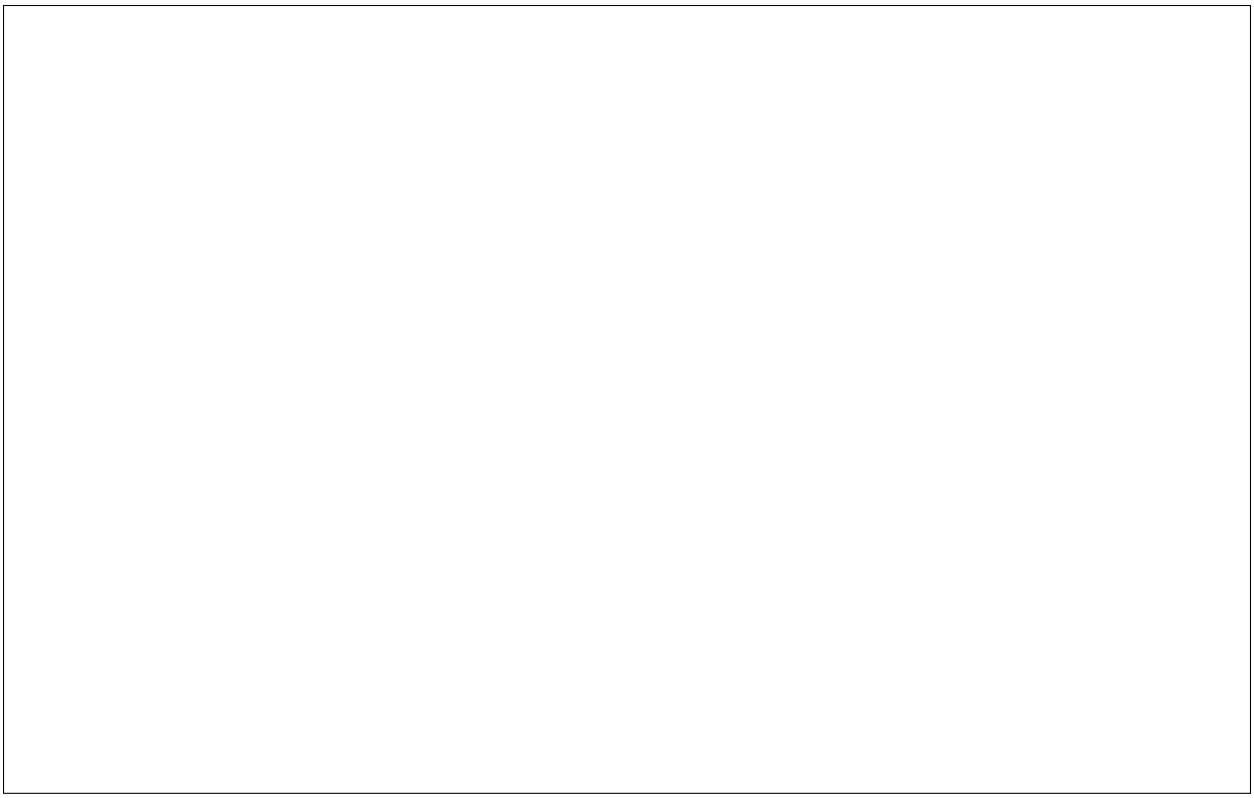
```
. reg logwage edu exper tenure married black south urban blackeduc
```

Source	SS	df	MS		
Model	42.0055468	8	5.25069335	Number of obs =	935
Residual	123.650736	926	.133532113	F(8, 926) =	39.32
				Prob > F =	0.0000
Total	165.656283	934	.177362188	R-squared =	0.2536
				Adj R-squared =	0.2471
				Root MSE =	.36542

logwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0671153	.0064277	10.44	0.000	.0545008	.0797298
exper	.0138259	.0031906	4.33	0.000	.0075642	.0200876
tenure	.011787	.0024529	4.81	0.000	.0069732	.0166009
married	.1989077	.0390474	5.09	0.000	.1222761	.2755393
black	.0948086	.2553994	0.37	0.711	-.4064202	.5960375
south	-.0894495	.0262769	-3.40	0.001	-.1410187	-.0378803
urban	.1838523	.0269547	6.82	0.000	.130953	.2367516
blackeduc	-.0226236	.0201827	-1.12	0.263	-.0622326	.0169854
_cons	5.374817	.1147027	46.86	0.000	5.14971	5.599925

Note: blackeduc = black*educ

5.2 (6 points) Interpret the coefficient of blackeduc. Set the hypothesis testing and test whether the return to education does depend on race at the 1 percent level of significance.



Again, start with the original model, but now allow wages to differ across four groups of people: married and black, married and nonblack, single and black, and single and nonblack.

Let us choose the base group to be single, nonblack. Then we add dummy variables: marrnonblck, singblck, and marrblck for the other three groups. The result is

$$\begin{array}{lllllll} \widehat{\log(wage)} = & 5.40 & +.0655 & educ+ & .0141 & exper+ & .0117 \quad tenure \\ & (.11) & (.0063) & & (.0032) & & (.0025) \\ & & -.092 & south+ & .184 & urban+ & .189 \quad marrnonblck \\ & & (.026) & & (.027) & & (.043) \\ & & -.241 & singblck+ & .0094 & marrblck & \\ & & (.096) & & (.0560) & & \end{array}$$

$$n = 935, R^2 = .253.$$

5.3 (6 points) What is the estimated wage differential between married blacks and married nonblacks?

———————————————The End of Exam———————————————