There are 6 questions. If you get stuck on one part, move on and do the rest. GOOD LUCK!

1. A few years ago, *New York Time*s published an article titled "Wine for the Heart: Over All, Risks May Outweigh Benefits."

Motivated by the article, we wish to estimate equations such as

$$heart = \beta_0 + \beta_{alc} alcohol + u.$$

The variables are
alcohol: per capita consumption of liters of wine
heart: deaths due to heart disease per 100,000

a. Say, we first estimate the following simple regression using data on $n = 150$ countries:

$$\widehat{heart} = 239.147 - 19.683 alcohol.$$

Interpret the slope in this equation and explain its sign and magnitude.

Answer: As per capita alcohol consumption increases by 1 liter, deaths due to heart disease decreases by 19.68 per 100,000 people.

b. Next, the following simple regression is also estimated:

$$\log(\widehat{heart}) = 5.361 - 0.353 \log(alcohol).$$

Interpret the slope in this equation and explain its sign and magnitude.

Answer: As per capita alcohol consumption increases by 1%, deaths due to heart disease per 100,000 people decreases by 0.353%.

c. Do the simple regressions above obtain unbiased estimators of the effect of country-level alcohol consumption and deaths due to heart disease? Explain.

Answer: The estimators above based on a simple regression model are unlikely to be unbiased. A number of factors such as a country's average education and income levels are likely correlated with alcohol consumption as well as heart disease.

2. The results below correspond to a regression output in Stata. The data set contains data on colleges and the variables are
enroll: total enrollment
police: employed officers
crime: total campus crimes
lcrime: log(crime)
lenroll: log(enroll)
lpolice: log(police).

The equation of interest is given by:

$$\log(crime) = \beta_0 + \beta_{lpolice} \log(police) + \beta_{lenroll} \log(enroll) + u.$$

```
. reg lcrime lpolice lenroll

      Source |       SS       df       MS              Number of obs =      97
-------------+------------------------------           F(  2,    94) =   80.72
       Model |  115.732673     2  57.8663366           Prob > F      =  0.0000
    Residual |  67.3868053    94  .716880908           R-squared     =  0.6320
-------------+------------------------------           Adj R-squared =  0.6242
       Total |  183.119479    96  1.90749457           Root MSE      =  .84669

------------------------------------------------------------------------------
      lcrime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lpolice |   .5163558   .1486583     3.47   0.001     .2211913    .8115203
     lenroll |   .9234745   .1439888     6.41   0.000     .6375813    1.209368
       _cons |  -4.793824   1.112043    -4.31   0.000     -7.00181   -2.585837
------------------------------------------------------------------------------
```

a. What does the R-squared value of 0.632 imply?

Answer: The R^2 value implies that about 63% of the variation in log(*crime*) is explained by log(*police*) and log(*enroll*).

b. Assuming a two-tailed test where H_0: $\beta_{lenroll} = 0$, is the coefficient estimate corresponding to log(*enroll*) statistically significant (i.e., H_0 is rejected) at the 2% level of significance?

Answer: Yes. For example, the p-value is less than 0.02.

c. What is the (numerical) value of the *t* test statistic to test whether the coefficient estimate corresponding to log(*police*) is significantly different from 0.7 (i.e., H_0: $\beta_{lpolice}=0.7$)? It is fine to use values up to two decimal points.

Answer: The test statistic is given by $(0.516 - 0.7)/0.149 = -1.23$.

d. Suppose we are jointly testing whether the slope coefficients corresponding to log(*police*) and log(*enroll*) are zero, i.e., H_0: $\beta_{\text{lenroll}}=0$ and $\beta_{\text{lpolice}}=0$. Write the R-squared form of *this* F statistic. Do you reject the hypothesis at the 1% level of significance?

Answer: In this case, the unrestricted R^2 is the R^2 from the above regression. The restricted R^2 is zero. So, the F statistic is given by $[R^2/q] / [(1 - R^2)/(n - k - 1)]$, i.e., $[0.632/2] / [(1 - 0.632)/(97 - 2 - 1)] = 80.72$. This is actually displayed at the top right hand corner of the results table. From Table G.3c, the critical value is about 4.85. Hence, we reject H_0.

3. Answer the following briefly:

a. In a regression model, what is the average (numerical) value of the residuals?

Answer: Zero.

b. In a regression model, what is the (numerical) value of correlation between the residuals and each explanatory variable?

Answer: Zero.

c. Is the assumption of homoskedasticity required for unbiasedness of $\hat{\beta}_j$?

Answer: No.

d. Is the assumption of normality required for unbiasedness of $\hat{\beta}_j$?

Answer: No.

4. Using cross section data on individuals in a certain year, the following equation is estimated

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$$

where the variables are
wage: average hourly earnings
educ: years of education
exper: years of experience.

The regression results are:

```
. reg wage educ exper expersq

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(3, 522)       =     64.11
       Model |  1927.87673         3  642.625576   Prob > F        =    0.0000
    Residual |  5232.53756       522  10.0240183   R-squared       =    0.2692
-------------+----------------------------------   Adj R-squared   =    0.2650
       Total |  7160.41429       525  13.6388844   Root MSE        =    3.1661

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5953429   .0530251    11.23   0.000     .4911741    .6995118
       exper |    .268287   .0368969     7.27   0.000     .1958023    .3407717
     expersq |  -.0046123    .000822    -5.61   0.000     -.006227   -.0029975
       _cons |   -3.96489   .7521526    -5.27   0.000    -5.442508   -2.487272
------------------------------------------------------------------------------
```

a. Using the formula for the effect of experience on wage in this setup, what is the return to the fifth year of experience, i.e., when *exper* increases from 4 to 5? Provide the numerical value.

Answer: The return to experience is given by $\beta_2 + 2\beta_3 exper$. For the fifth year of experience, this is 0.268 + 2(-0.0046)4, i.e., 0.231.

b. At what value of *exper* does additional experience actually begin to lower predicted wage (i.e., the turning point)? Provide the numerical value.

Answer: The turning point is given by |0.268/2(-0.005)|, i.e., 26.8 years.

5. Consider a model where the return to education depends upon the amount of work experience (and vice versa):

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 educ \cdot exper + u$$

where the variables and data are as discussed in question 4.

The summary statistics are

```
. su wage educ exper
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	526	5.896103	3.693086	.53	24.98
educ	526	12.56274	2.769022	0	18
exper	526	17.01711	13.57216	1	51

The regression results are

```
. g educexper = educ*exper

. reg wage educ exper educexper
```

Source	SS	df	MS
Model	1615.96222	3	538.654074
Residual	5544.45207	522	10.6215557
Total	7160.41429	525	13.6388844

Number of obs	= 526
F(3, 522)	= 50.71
Prob > F	= 0.0000
R-squared	= 0.2257
Adj R-squared	= 0.2212
Root MSE	= 3.2591

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.6017355	.0899	6.69	0.000	.4251253	.7783457
exper	.0457689	.0426138	1.07	0.283	-.0379466	.1294844
educexper	.0020623	.0034906	0.59	0.555	-.004795	.0089197
_cons	-2.859916	1.18108	-2.42	0.016	-5.180169	-.5396623

a. Using the formula for the effect of experience on wage in this setup, what is the approximate return to experience for the average level of education? Provide the numerical value.

Answer: The return to experience is given by $\beta_2 + \beta_3 educ$. For the average level of education, this is 0.046 + 0.002 x 12.56, i.e., 0.071.

b. Using the formula for the effect of education on wage in this setup, what is the approximate return to education for the average level of experience? Provide the numerical value.

Answer: The return to education is given by $\beta_1 + \beta_3 exper$. For the average level of experience, this is 0.602 + 0.002 x 17.02, i.e., 0.636.

6. Consider a data set on prices of homes sold in the Boston, MA area around 1990. The variables are
price: house price, $1000s
bdrms: number of bdrms
sqrft: size of house in square feet.

Suppose that the multiple regression of log(*price*) on *bdrms* and log(*sqrft*) satisfies the assumptions required for unbiasedness. However, the simple regression of log(*price*) on *bdrms* does not. In fact, the coefficient estimate of *bdrms* likely suffers from an omitted variable bias in case of the simple regression. If log(*sqrft*) has a positive effect on log(*price*) and log(*sqrft*) and *bdrms* are positively correlated, what is the likely sign of the omitted variable bias?

Answer: If the omitted variable log(*sqrft*) has a positive effect on log(*price*), and log(*sqrft*) and *bdrms* are positively correlated, the likely sign of the omitted variable bias is positive. In other words, the coefficient estimate corresponding to *bdrms* in case of the simple regression of log(*price*) on *bdrms* is likely upward biased.