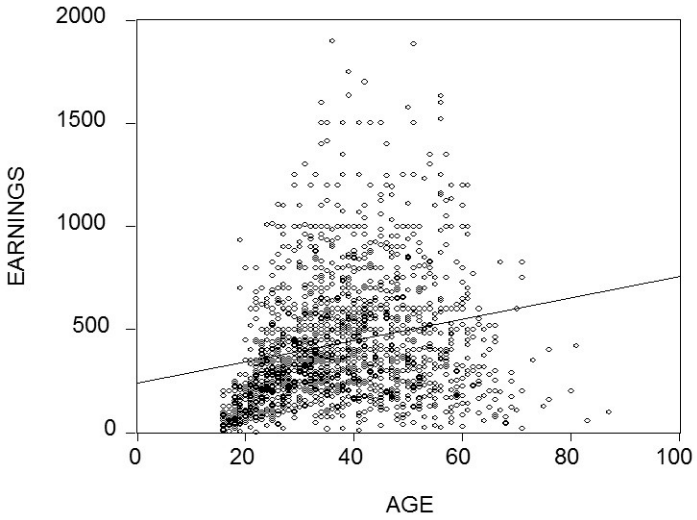
Problem set 3

Due on Thursday October 6, 2022 (by 11:59 PM EST)

Note: No credit will be given if you report only the final answers without showing formulas and calculations when appropriate. This applies to both theoretical and empirical questions. For the empirical questions, make sure to upload the R scripts and output through Latte before the deadline. No credit will be given if the R output is missing.

Problem 1

After analyzing the age-earnings profile for 1,744 workers as shown in the figure below, it becomes clear to you that the relationship cannot be approximately linear.



You estimate the following polynomial regression model, controlling for the effect of gender by using a binary variable that takes on the value of one for females and is zero otherwise:

$$\widehat{Earn} = \underset{(283.11)}{-795.90} + \underset{(29.29)}{82.93} \times Age - \underset{(1.06)}{1.69} \times Age^2 + \underset{(0.016)}{0.015} \times Age^3 - \underset{(0.0009)}{0.0005} \times Age^4$$

$$- \underset{(12.45)}{163.19}\, Female,\ R^2 = 0.225,\ SER = 259.78$$

(a) Test for the significance of the Age^4 coefficient. Describe the general strategy to determine the appropriate degree of the polynomial.

(b) You run two further regressions. Present an argument as to which one you should use for further analysis.

$$\widehat{Earn} = \underset{(120.13)}{-683.21} + \underset{(9.27)}{65.83} \times Age - \underset{(0.22)}{1.05} \times Age^2 + \underset{(0.002)}{0.005} \times Age^3$$

$$\underset{(12.45)}{-163.23}\ Female,\ R^2 = 0.225,\ SER = 259.73$$

$$\widehat{Earn} = \underset{(51.58)}{-344.88} + \underset{(2.64)}{41.48} \times Age - \underset{(0.03)}{0.45} \times Age^2$$

$$\underset{(12.47)}{-163.81}\ Female,\ R^2 = 0.222,\ SER = 260.22$$

(c) Sketch the graph of fitted earnings of males against age of your preferred regression. Does this make sense? Are you concerned about the negative coefficient on the regression intercept? What is the implication for female earners in this sample?

(d) Explain how you would calculate the effect of changing age by one year on earnings, holding constant the gender variable. Finally, briefly describe how you would calculate the standard errors of the estimated effect.

Problem 2

You have estimated an earnings function, where you regressed the log of earnings on a set of continuous explanatory variables (in levels) and two binary variables, one for gender and the other for marital status. One of the explanatory variables is education.

(a) Interpret the education coefficient.

(b) Next, specify the binary variables and an equation, where the default is a single male, without allowing for interaction between marital status and gender. Indicate the coefficients that measure the effect of a single male, single female, married male, and married female.

(c) Finally allow for an interaction between the gender and marital status binary variables. Repeat the exercise of writing down the various effects based on the female/male and single/married status. Why is the latter approach more general than the former?

Problem 3

Several authors have tried to measure the "persistence" in U.S state unemployment rates by running the following regression:

$$ur_{i,t} = \beta_0 + \beta_1 \times ur_{i,t-k} + z_{i,t}$$

where *ur* is the state unemployment rate, *i* is the index for the *i*-th state, *t* indicates a time period, and typically $k \geq 10$.

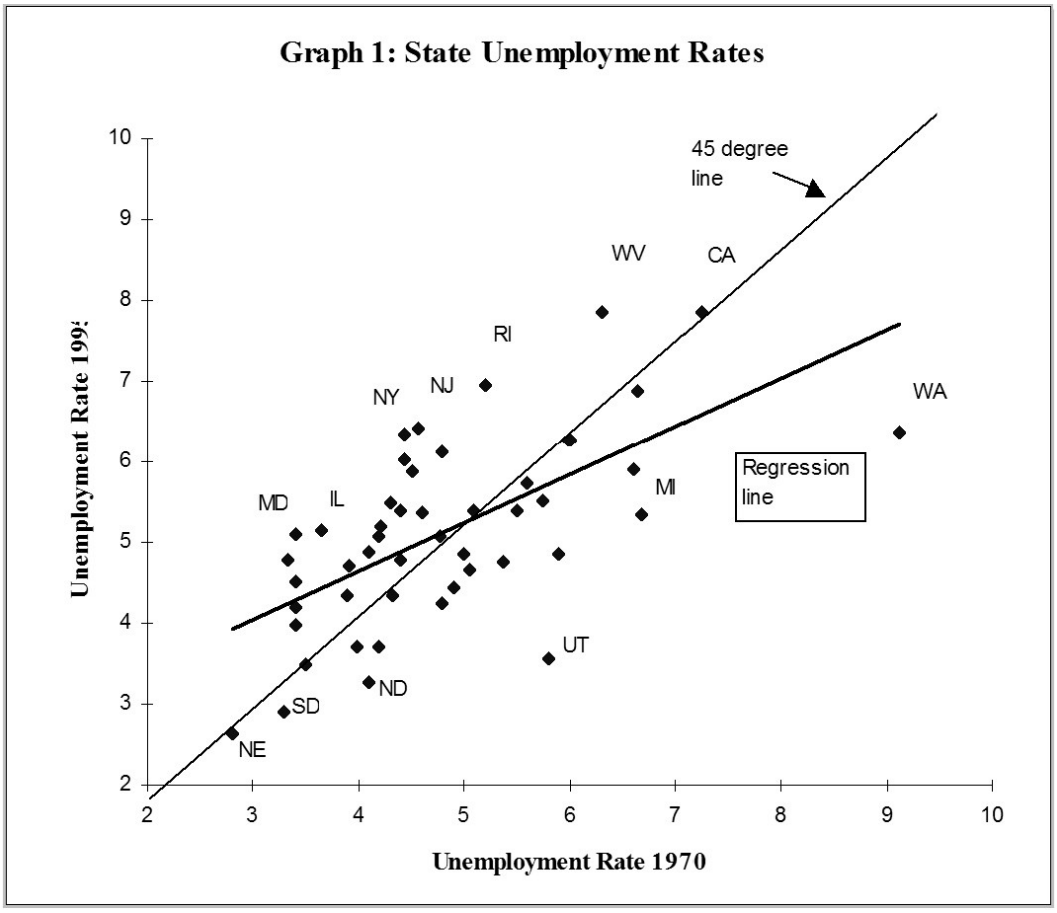
(a) Explain why finding a slope estimate of one and an intercept of zero is typically interpreted as evidence of "persistence."

(b) You collect data on the 48 contiguous U.S. states' unemployment rates and find the following estimates:

$$\hat{ur}_{i,1995} = 2.25 + 0.60 \times ur_{i,1970};\ R^2 = 0.40,\ SER = 0.90$$
$$(0.61)\ \ (0.13)$$

Interpret the regression results.

(c) Analyze the accompanying figure, and interpret the observation for Maryland (MD) and for Washington (WA). Do you find evidence of persistence? How would you test for it?



(d) One of your peers points out that the results from the regression above makes little sense, since it implies that eventually all states would have identical unemployment rates. Explain the argument.

(e) Imagine that state unemployment rates were determined by their natural rates and some transitory shock. The natural rates themselves may be functions of the unemployment insurance benefits of the state, unionization rates of its labor force, demographics, sectoral composition, etc. The transitory components may include state-specific shocks to its terms of trade such as raw material movements and demand shocks from the other states. You specify the *i*-th state unemployment rate accordingly as follows for the two periods when you observe it,

$$\tilde{X}_{i,t} = X_i + v_{i,t} \text{ and } \tilde{X}_{i,t-k} = X_i + w_{i,t-k},$$

so that actual unemployment rates are measured with error. You have also assumed that the natural rate is the same for both periods. Subtracting the second period from the first then results in the following population regression function:

$$\tilde{X}_{i,t} = 0 + 1 \times \tilde{X}_{i,t-k} + (v_{i,t} - w_{i,t-k})$$

With this setup, it can be shown that estimation of the observed unemployment rate in period *t* on the unemployment rate in period (*t-k*) by OLS results in an estimator for the slope coefficient that is biased towards zero. The formula showing this is,

$$\hat{\beta}_1 \xrightarrow{p} 1 - \frac{\sigma_v^2}{\sigma_X^2 + \sigma_v^2}.$$

Where σ_X^2 denotes the variance of the natural rate, while σ_v^2 represents the variance of the transitory part. Using this insight, explain over which periods you would expect the slope to be closer to one, and over which period it should be closer to zero.

(f) Estimating the same regression for a different time period results in

$$\hat{ur}_{i,1995} = 3.19 + 0.27 \times ur_{i,1985}; \; R^2 = 0.21, \; SER = 1.03$$
$$(0.56) \; (0.07)$$

If your above analysis is correct, what are the implications for this time period?

Problem 4 (empirical)

What are the root causes of terrorism? Poverty? Repressive political regimes? Religious or ethnic conflicts arising from heterogenous populations? In this problem you will examine empirical evidence on cross-country sources of terrorism. Variables in the data set terrorism.dta, are defined at the end of the problem set. The data set is posted on Latte.

To do this problem, you will need to create (generate) some new variables, which are functions of the variables in terrorism.dta. You should restrict your analysis to the observations for which there are data on GDP per capita (i.e. for which gdppc is non-missing) and for which there are terrorist fatalities (i.e. for which ftmpop > 0).

a. Produce the scatterplot of ftmpop vs gdppc.

b. Generate the variables lnftmpop = log(ftmpop) and lngdppc = log(gdppc). Produce the scatterplot of lnftmpop vs. lngdppc.

c. Using the scatterplots from (a) and (b), would you suggest the variables (i) ftmpop and gdppc or (ii) lnftmpop and lngdppc for modeling using OLS?

d. Estimate the following three regressions and attach the R output to your problem set:

(i) Regression 1: The natural logarithm of *ftmpop* (the dependent variable) regressed on *lngdppc*, *lackpf* , $lackpf^2$ (i.e., *lackpf* squared), *ethnic*, and *religion*.

(ii) Regression 2: The natural logarithm of *ftmpop* (the dependent variable) regressed on *lackpf*, *higdppc* and (*higdpcc x lackpf*)

(iii) Regression 3: The natural logarithm of *ftmpop* (the dependent variable) regressed on *lngdppc*, *lackpf* , $lackpf^2$, *higdppc*, *ethnic*, *religion*, (*higdpcc x ethnic*) and (*higdpcc x religion*).

Use these results to answer the following questions:

e. Using regression (1), estimate the effect on *lnftmpop* of changing from *lackpf* = 7 (extremely limited political freedoms) to *lackpf* = 5 (some political freedoms), holding constant the values of the other regressors in regression (1).

f. Regression (2) produces two regression lines, one for *higdppc* = 0 and one for *higdppc* = 1. Write out the estimated regression lines (in equation form) for the two groups. Is the difference between the two slopes statistically significantly different from zero at the 5% significance level? Explain.

g. Using regression (2), explain in words what the coefficient on the interaction term `higdpcc * lackpf` means.

h. One theory is that ethnic and religious diversity leads to strife and terrorism when economic resources are poor (i.e. GDP per capita is below the median), but if overall economic conditions are strong (GDP per capita above the median) then ethnic and religious diversity are more readily tolerated and do not lead to terrorism. Does the regression in regression (3) support or refute this theory? Explain and clearly describe the test you are using to test this hypothesis.

Data description: terrorism.dta

Variable	Definition
ftmpop	Number of fatalities from terrorist incidents in the country, 1998-2004, per million population (U.S. State Department)
gdppc	GDP per capita in the country (World Bank)
higdppc	Dummy variable = 0 if GDP per capita is below the median; = 1 if GDP per capita is above the median
lackpf	Index of the lack of political freedoms (Freedom House), 1 to 7 scale; 7 = extremely limited political freedoms
ethnic	Index of ethnic fractionalization (0 to 1 scale, 0 = no fractionalization)
religion	Index of religious fractionalization (0 to 1 scale, 0 = no fractionalization)