

**Advanced Regression  
Problem Set 3  
Due: November 16, 2009**

1. During the past decade, the state of Maryland has encountered several fish kills around the Chesapeake Bay. These fish kills have been linked primarily to algae called *pfiesteria psycicida*. Discovered in 1988, *pfiesteria* is a toxic dinoflagellate that has a complex life cycle and exists in 24 reported forms. *Pfiesteria* generally exists benignly, feeding on bacteria in the water and sediments of tidal waters. Occasionally, however, something excreted by schools of fish triggers a *pfiesteria* outbreak in waters. When this triggering event occurs, *pfiesteria* begins to release a toxin into the water, which causes fish to become less active. It is thought that other toxins produced by the dinoflagellate break down fish skin tissue and adversely alter the internal salt balance of fish, causing lesions. As the fish become injured, the predatory *pfiesteria* feeds on their tissues and blood, causing death through secondary infections. Typically, an outbreak will last only a few hours before the dinoflagellates change back into a non-toxic form, but its effects may be seen in resident fish for the next several days or weeks.

In response to an increasing number of fish kills, the state of Maryland in 1987 (working with the state of Virginia, the District of Columbia, and the Environmental Protection Agency) entered into a regional agreement to work to reduce pollution from point sources in the Chesapeake Bay region. This agreement took effect in 1992, and was directed at factories, boats on the Chesapeake, and other point sources of pollution along bay tributaries.

*Pfiesteria* outbreaks are typically triggered by dense fish populations, but can be triggered by other water quality factors. Specifically, measurable factors such as

- dissolved oxygen (DO),
- total nitrogen (TN),
- total phosphorous (TP), and
- total suspended solids (TSS)

have been shown to be related to *pfiesteria* outbreaks.

Low levels of DO in warmer parts of the year precede *pfiesteria* fish kills. Levels of TN and TP are common nutrients in the Chesapeake Bay area, and human development has caused increased flows of nutrients into the Bay. It is believed that higher levels of TN and TP are associated with a greater likelihood of observing a *pfiesteria* outbreak. TSS is a measure of siltation into the Bay, also influenced by increased human development; higher levels of TSS have been associated with fish kills as well.

Data for this analysis comes from the Water Quality Monitoring Program of the Chesapeake Bay, which began in 1984. Data come from 53 stations in Maryland over the years 1990 to 1999. Data are collected monthly in colder months and biweekly in warmer months.

**Table 1.1**  
**Pollution Levels Before and After 1987 Agreements Went Into Effect (in 1992)**

	<b>Pre-1992</b>	<b>Post-1992</b>
Dissolved Oxygen (DO)	7.900 (3.067) [4,185]	7.881 (3.127) [13,820]
Total Nitrogen (TN)	0.899 (0.636) [4,161]	0.875 (0.501) [13,787]
Total Phosphorous (TP)	0.060 (0.058) [4,082]	0.060 (0.053) [13,789]
Total Suspended Solids (TSS)	16.356 (18.234) [4,184]	16.149 (21.547) [13,818]

*Note:* All pollutants are measured in mg per liter. Standard deviations are in parentheses, and sample sizes are in brackets.  
*Source:* The Water Quality Monitoring Program of the Chesapeake Bay, 1990 to 1999.

a) Table 1.1 contains information on pre- and post-pollution levels around the implementation of the 1987 agreement in 1992. Is there a statistically significant difference in pollution levels pre- and post-1992 for DO? TSS?

**Table 1.2**  
**Regression Results for DO, TN, TP, and TSS**

	<b>Dependent Variable</b>			
	<b>DO</b> (1) DO in mg/L	<b>TN</b> (2) ln(tn)	<b>TP</b> (3) ln(tp)	<b>TSS</b> (4) ln(tss)
1987 Agreement in Effect (1992 on)	-0.871 (10.57)**	-0.257 (33.87)**	-0.047 (3.21)**	-4.222 (92.85)**
Other Controls Included	Yes	Yes	Yes	Yes
Station Fixed Effects	Yes	Yes	Yes	Yes
Year Dummies	Yes	Yes	Yes	Yes
Observations	17,991	17,934	17,857	17,988
R-squared	0.57	0.76	0.73	0.73

*Note:* t-statistics are in parentheses. \* indicates statistical significance at the 5% level, and \*\* indicates statistical significance at the 1% level. Other controls include a measure of water temperature at the time of the reading and the depth of the reading. Station fixed effects include 52 station-specific dummy variables. Year fixed effects include 9 year-specific dummy variables.

The results presented in Table 1.2 come from a fixed effects regression of each dependent variable (DO, TN, TP, TSS) on a dummy variable that equals unity for all station-years on/after 1992, other control variables, station fixed effects, and year dummy variables.

b) In this fixed effects regression, what is the unit of analysis?

c) Has the 1987 agreement had the intended effect on pollution emissions? Interpret the coefficient on the 1987 agreement dummy variable for each model.

*d)* All models contain station and year fixed effects. Explain what each fixed effect is capturing. Be clear in your description.

2. The Earned Income Tax Credit (EITC) is now a major piece of U.S. anti-poverty policy. Enacted in 1975, the EITC provides a refundable tax credit to low-income working families that offsets both federal/state income taxes and the federal payroll tax. If a family's total tax liability is less than the amount of the EITC, the family receives the difference in the form of a tax refund from the IRS.

In order to qualify for the EITC, individuals must be working, have earnings below an eligibility threshold (which varies by state, year, and number of children), and have at least one child ages 0 to 18 (although the EITC provides small credits for single individuals). The qualifying child must be a child, grandchild, stepchild, or foster child of the taxpayer, and lives with the taxpayer for the entire tax year.

As of 2003, a taxpayer with earnings and one qualifying child is eligible to receive a 34 percent wage subsidy on earnings up to \$7,490, for a maximum credit of \$2,547. Taxpayers with two or more children receive a 40 percent subsidy on earnings up to \$10,510, for a maximum credit of \$4,204.

Not surprisingly, the EITC has been hailed as an effective method for moving unemployed welfare recipients off cash assistance and into work. However, this policy may also distort marriage behavior in ways that are not desirable from the perspective of policymakers. For example, if an employed single mother who is eligible for the EITC marries another low-wage worker, their combined earnings could make them ineligible for the credit. This may reduce the incentive to become married.

In this problem, we will take a closer look at the impact of the EITC on marriage. Specifically, we will use a panel dataset of all states (and the District of Columbia) over the years 1977 to 2004 to explore the relationship between the EITC maximum credit and a measure of new marriage rates. Because our data vary across states and over time, we will make use of state and year fixed effects in our regressions. Use the dataset called *eitc\_marriage* for your analyses.

*a)* To familiarize yourself with the dataset, begin by describing the data and producing some summary statistics. What is the unit of analysis?

*b)* Estimate a simple OLS regression of  $\ln(\text{new marriage rate})$  on the EITC maximum credit. Be sure to estimate all models with robust standard errors. Interpret the coefficient on *eitc*. Is the coefficient statistically significant? If so, at what level?

*c)* Now, add the remaining "observables" to the model (i.e., the rest of the variables in the dataset). Why did I refer to these variables as "observables?" What happens to the coefficient on *eitc*? Does this make sense? What happens to the  $R^2$ ? Does this make sense?

*d)* Next, estimate a fixed effects regression, adding state and year fixed effects to the model (in addition to all other variables). To estimate this model, type the following:<sup>1</sup>

```
xi: regress lnmarry eitc waiver tanf welf ur lnemp_s_r earn_s_r pdens black  
fem_15_39 fem_40_64 ba rgov rpres i.fips i.year, robust
```

---

<sup>1</sup> To get help with this Stata code type `help xi`.

What do the `i.fips` and `i.year` indicate? Look carefully at the model's output. What are the variables below the observable state-year characteristics? Hint: Look at the variable list, and tabulate a few of the newly created variables. What are these?

*e)* At the top of Stata's output for this model, you should see the following:

```
i.fips      _Ifips_1-56      (naturally coded; _Ifips_1 omitted)
i.year      _Iyear_1977-2004 (naturally coded; _Iyear_1977 omitted)
```

Describe what this means. Now, interpret the coefficient on `_Iyear_1978`. Interpret the coefficient on `_Iyear_2004`. What does this tell you about marriage rates between 1977 and 2004?

Explain clearly what the state and year fixed effects are capturing.

*f)* Interpret the coefficient on `eitc` for the fixed effects regression. Describe any changes from the previous model. What does this tell you about the earlier models? Specifically, in the absence of fixed effects, is there any reason to believe that the earlier results are biased?