Problem 2

A. Categorical variable: *no_toilet*

0 "Has toilet" (value "label")	1 "No toilet" (value "label")
10 flush toilet 11 flush - to piped sewer system 12 flush - to septic tank 13 flush - to pit latrine 14 flush - to somewhere else 15 flush - don't know where 20 pit toilet latrine 21 pit latrine - ventilated improved pit (vip) 22 pit latrine - with slab 41 composting toilet	23 pit latrine - without slab / open pit 30 no facility 31 no facility/bush/field 42 bucket toilet 43 hanging toilet/latrine 96 other

Categorical variable: *poor_water*

0 "Good Water Resource" (value "label")	1 "Poor Water Resource" (value "label")
10 piped water 11 piped into dwelling 12 piped to yard/plot 13 public tap/standpipe 14 piped to neighbor 20 tube well water 21 tube well or borehole 30 dug well (open/protected) 31 protected well 33 protected public well 41 protected spring 61 tanker truck 62 cart with small tank 71 bottled water	22 open well in yard 23 open public well 24 open well neighbor 32 unprotected well 40 surface water 42 unprotected spring 43 river/dam/lake/ponds/stream/canal /irrigation channel 51 rainwater 96 other

B.

Year	Child had fever in last two weeks		Difference (a-b)
	0 (a)	1 (b)	
2001 (A)	0.5541	0.4459	0.1082
2007 (B)	0.821	0.179	0.642
Difference (B-A)	0.2669	-0.2669	

We can see that compared to 2001, the probability of a child having a fever in the last two weeks has decreased by about 27%.

. In 2001, about 11% less children had a fever in the last two weeks than didn't while in 2007, about 64% less children had a fever in the last two weeks than didn't. The trend is that the probability of a child having a fever in the last two weeks has reduced.

C.

```
Linear regression                               Number of obs     =     11,536
                                                F(1, 71)          =       0.11
                                                Prob > F          =     0.7394
                                                R-squared         =     0.0000
                                                Root MSE          =     .46435

                                   (Std. Err. adjusted for 72 clusters in NBER_District_ID)
------------------------------------------------------------------------------
             |               Robust
       fever |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
treatment_d~t |  -.0064191    .01922    -0.33   0.739    -.0447426    .0319044
       _cons |   .3155755   .010579    29.83   0.000     .2944815    .3366694
------------------------------------------------------------------------------
```

The coefficient on the *treatment_district* is -0.00642. In this specification, compared to an untreated districted, a child in treated district has 0.64% less chance of having a fever in the last two weeks. However, this coefficient is not significant at the 5% or 10% significance level.

D.

Linear regression

Number of obs	= 11,536
F(2, 71)	= 156.35
Prob > F	= 0.0000
R-squared	= 0.0829
Root MSE	= .44472

(Std. Err. adjusted for 72 clusters in NBER_District_ID)

fever	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
treatment_district	.0220845	.0148856	1.48	0.142	-.0075966	.0517656
post	-.2681343	.0154463	-17.36	0.000	-.2989333	-.2373353
_cons	.4427939	.014291	30.98	0.000	.4142984	.4712894

The coefficient on *post* is -0.268. In this new specification, keeping the treatment assignment the same, in 2007 a child has 27% less chance of having a fever in the last weeks compared to 2001.

The coefficient on the *treatment_district* is 0.0221. It has increased (even changed sign) compared to its coefficient (-0.00642) in the previous specification (in c). What this means that keeping *post* the same, a child in the treatment district is 2.2% more likely to have a fever than a child in the untreated district. Even though I expected the coefficient to become less negative (due to correction from Omitted Variable Bias), I didn't expect it to be positive. This seems counterintuitive. But again, this coefficient is not significant at 5% or 10% significance level.

This is not the DID estimate.

E.

Linear regression

Number of obs	= 11,536
F(3, 71)	= 180.11
Prob > F	= 0.0000
R-squared	= 0.0863
Root MSE	= .44391

(Std. Err. adjusted for 72 clusters in NBER_District_ID)

fever	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
treatment_district	.1049327	.0226838	4.63	0.000	.0597026	.1501628
post	-.2438323	.0148573	-16.41	0.000	-.2734569	-.2142077
inter	-.1471027	.0285866	-5.15	0.000	-.2041029	-.0901026
_cons	.4312636	.0142101	30.35	0.000	.4029295	.4595978

α, the constant, is 0.4313: In this specification, in an untreated district in the year 2001, a child had a 43% chance of having a fever in the last two weeks.

β, the coefficient on *treatment_district*, is 0.105: In this specification, in the year 2001, compared to an untreated district, a child in a treated district has 10.5% higher probability of having a fever in the last two weeks.

γ, the coefficient on *post*, is -0.244: In this specification, in an untreated district, compared to 2001, a child has 24% lesser probability of having a fever in the year 2007.

δ, the coefficient on *post x treatment_district*, is -0.147: In this specification, a child in a treated district in the year 2007 is 14.7% less likely to have a fever in the last two weeks compared to a child in an untreated district. Similarly, in 2007, a child in a treated district is 14.7% less likely to have a fever in the last two weeks compared to 2001.

This is the correct DID regression. The coefficient on *post x treatment_district, δ,* is the DID coefficient (estimated treatment effect).

F.

Linear regression

Number of obs	= 11,203
F(11, 71)	= 128.58
Prob > F	= 0.0000
R-squared	= 0.1100
Root MSE	= .43796

(Std. Err. adjusted for 72 clusters in NBER_District_ID)

fever	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
treatment_district	.0972236	.0210181	4.63	0.000	.0553147	.1391326
post	-.2445625	.0154882	-15.79	0.000	-.2754452	-.2136799
inter	-.1414374	.0281241	-5.03	0.000	-.1975152	-.0853596
no_toilet	.0118389	.0098729	1.20	0.234	-.0078472	.0315249
poor_water	.0154237	.0145666	1.06	0.293	-.0136213	.0444688
age						
1	.0696039	.0122057	5.70	0.000	.0452665	.0939414
2	.0152087	.0126778	1.20	0.234	-.0100701	.0404875
3	-.0719995	.0113925	-6.32	0.000	-.0947156	-.0492834
4	-.1192792	.0113248	-10.53	0.000	-.1418602	-.0966982
v133	-.0067443	.0018788	-3.59	0.001	-.0104906	-.002998
v136	.0017273	.0019738	0.88	0.384	-.0022082	.0056629
_cons	.4572427	.0244049	18.74	0.000	.4085807	.5059047

The coefficient on *no_toilet* is -0.01184: In this model specification, everything else being constant, a child living in a household without a toilet is 1.184% less likely to have a fever compared to a child living in a household with a toilet.

The coefficient on *poor_water* is 0.015424: In this model specification, everything else being constant, a child living in a household with a poor water source is 1.54% more likely to have a fever compared to a child living in a household with a good water resource.

The coefficient on 1.*age* is 0.0696: In this model specification, everything else being constant, a child in age group 1 is, on average, associated with 6.96% higher probability of having a fever in the next two weeks, compared to a child in age group 0.

The coefficient on 2.*age* is 0.01521: In this model specification, everything else being constant, a child of age group 2 is, on average, associated with 1.52% higher probability of having a fever in the next two weeks, compared to a child in age group 0. However, this coefficient is not significant at the 5% or 10% significance level.

The coefficient on 3.*age* is -0.072: In this model specification, everything else being constant, a child of age group 3 is, on average, associated with 7.2% lower probability of having a fever in the next two weeks, compared to a child in age group 0.

The coefficient on 4.*age* is -0.1193: In this model specification, everything else being constant, a child of age group 4 is, on average, associated with 11.93% lower probability of having a fever in the next two weeks, compared to a child in age group 0.

The coefficient on *v133* is -0.006744: In this model specification, everything else being constant, an increase in the education years of the household by one year is, on average, associated with a 0.67% lower probability of the child having a fever.

The coefficient on *v136* is 0.00173: In this model specification, everything else being constant, an increase in the number of household members by one year is, on average, associated with a 0.173% higher probability of the child having a fever. However, this coefficient is not significant at the 5% or 10% significance level.

The DID coefficient now is -0.1414 and it is slightly smaller in magnitude compared to the previous coefficient (-0.147) in the model without the additional covariates. The DID coefficient is expected to change because variables such as nature of toilet facility, type of available water resource, age of the child, number and education of household members affect the chances of whether or not the child has a fever. However, if every household in the sample had access to same (or very similar) toilet and water resources, had same number of members and educated for the same number of years, then we would not see a change in the DID coefficient.

G. If a local NGO had run their own antimalarial campaign in districts NOT targeted by the national program, that would mean that the variable *treatment_district* should assume the value 1 for all observations (if their campaign is similar as the national campaign). So, the coefficient on the interaction term *post x treatment_district* should be the same as the coefficient on *post* (and we would have to remove the constant from the regression specification to ensure full column rank, i.e reg y x1...xn, noc). We should get DID coefficient as 0, i.e the same results as the following model:

$$fever_{it} = \alpha + \gamma . post_t + \epsilon_{it}$$

where α is the coefficient on *treatment_district* (which is 1 everywhere due to the parallel campaign run by the NGO)

However, we do get significant non-zero DID estimates despite our control group being contaminated. So, we might be underestimating the DID coefficient and the true estimate should actually have been higher.

H. If there was a nationwide drought in 2006/2007, and the mosquito population was unable to replicate at the usual rate, it would not affect the DID coefficient because it would affect the treatment and control districts similarly and hence any change in fever prevalence would get cancelled out when we take difference of the differences. So, any effect due to the nationwide drought would not matter in the difference in differences.

However, if there is some differential impact of the drought in different areas, then we would need to control for that effect in our regression specification. Otherwise, we might be wrongly estimating the DID coefficient.