Economics 172: Problem Set #3

Due on November 17, 2025 at 11:00pm

Zachary Brandt
zbrandt@berkeley.edu

Problem 1: Weather and Witch Killing

This question builds on the econometric analysis in *Miguel 2005* about Tanzania poverty and witch-killing. You will carry out some econometrics analysis related to that paper and also related to the *Miguel, Satyanath and Sergenti (2004)* article. You may write up your answers using a word processor, include copies/screen-shots of your regression tables, and attach a copy of your R script at the end. Alternatively, you may choose to produce an RMarkdown file that integrates your code, your written responses, and tables displaying your regression results into a single document (please "knit to PDF" and turn in the resulting PDF file; do not simply turn in your `.rmd` file).

In either case, your submission must include: Your entire R code/script; Your written answers; Your regression output. Please merge all documents into a single PDF before submitting.

Please download "pset3-2025-killing.csv" from the bCourses page.

Use the `read.csv` command to open it in R (or RStudio), either on your local computer, or on UC Berkeley's DataHub, found at https://r.datahub.berkeley.edu/.

This dataset is a partial extract of the data from *Miguel (2005)*, organized such that each observation (row) contains data for a particular village (denoted by the variable vid) in a particular year (denoted year) in Meatu district, Tanzania. In other words, this is panel data.

Variables include:

- `witch_murders`: number of witch murders in a given year village-year
- `oth_murders`: number of non-witch murders
- `any_rain`: indicator (1/0) for whether a drought or flood occurred
- `any_disease`: indicator for whether a disease outbreak (measles, cholera, etc.) occurred
- `famine`: indicator for whether there was an extreme food shortage
- `educat`: average years of schooling in the village
- `trad_relig`: proportion of households practicing tradiational religions

a) Construct a new variable for the total number of murders in a village-year (witch + non-witch murders).

b) Create a table of summary statistics for all variables in the dataset, including the mean, standard deviation, minimum, maximum, and number of observations, using `stargazer`, `summary` or `describe` commands in R. Discuss any noteworthy patterns. Pay particular attention to the murder and rainfall variables.

c) Now consider the effect of extreme weather on murders in the village.

   (i) Install "miceadds" and "sandwich." Using the `lm.cluster` command, regress total murders (in a village in a particular year) on the indicator for whether a drought or flood occurred in that year. Make sure that error terms should be allowed to be correlated ("clustered") across years for the same village (use `vid`). Simply use `summary` to report the results in this question. [Note: Results estimated by `lm.cluster` could not be exported directly with `stargazer` so we use `summary` for simplicity. In the section we will teach how to export clustered regression results in a neater way.]

   (ii) In a second regression, add average years of schooling and proportion of households practicing traditional religions as additional explanatory variables.

   (iii) Interpret both regressions carefully.

Problem 1 continued on next page...

d) Finally, consider a possible instrumental variables (IV) approach. Economic theory suggests that extreme economic hardship—such as a famine—may be associated with more violence, including murders. Famine may be caused by extreme rainfall (which would be the instrumental variable).

(i) Write out the first stage regression, the second stage regression, and the reduced form regression.

(ii) Evaluate whether this is a valid IV approach by discussing the plausibility of the three key IV conditions: Relevance; Exclusion restriction; Exogeneity. What are some specific ways in which each of these assumptions might be appropriate or might fail in this context?

Solution

a) The new variable for total murders is created in R as follows:

```
killing$total_murders <- killing$witch_murders + killing$oth_murders
```

b) Summary statistics for all variables in the dataset are shown below.

Statistic	N	Mean	St. Dev.	Min	Max
vid	736	35.034	20.660	1	71
year	736	1,996.993	3.161	1,992	2,002
witch_murders	736	0.091	0.323	0	3
oth_murders	736	0.091	0.395	0	5
any_rain	736	0.171	0.377	0	1
any_disease	736	0.148	0.355	0	1
famine	736	0.174	0.379	0	1
educat	736	4.035	1.068	0.857	6.667
trad_relig	736	0.654	0.206	0.000	1.000
total_murders	736	0.182	0.516	0	5

The mean number of witch murders (0.091) and other murders (0.091) are identical. Droughts or floods occur in approximately 17% of observations, while extreme food shortages occur in 17.4% of village-years. The strong correlation between these suggests that weather shocks may be driving food insecurity. On average, villagers have about 4 years of schooling, and 65% of households practice traditional religions. The murder variables are right-skewed, with most village-years experiencing zero murders and maximum values of 3 for witch murders and 5 for other murders.

c) (i) The first regression estimates the effect of extreme weather on total murders with clustered standard errors:

$$\text{total\_murders}_{it} = \beta_0 + \beta_1 \cdot \text{any\_rain}_{it} + \epsilon_{it}$$

The estimated regression equation is

$$\widehat{\text{total\_murders}}_{it} = 0.174 + 0.048 \cdot \text{any\_rain}_{it}$$

where standard errors (clustered by village) are 0.022 for the constant and 0.046 for any_rain. The coefficient is not statistically significant ($p = 0.289$), and the $R^2 = 0.00125$ indicates that extreme rainfall explains very little variation in murders.

(ii) Adding education and traditional religion controls:

$$\text{total\_murders}_{it} = \beta_0 + \beta_1 \cdot \text{any\_rain}_{it} + \beta_2 \cdot \text{educat}_{it} + \beta_3 \cdot \text{trad\_relig}_{it} + \epsilon_{it}$$

Problem 1 continued on next page...

The estimated regression equation is:

$$\widehat{\text{total\_murders}}_{it} = 0.328 + 0.040 \cdot \text{any\_rain}_{it} - 0.038 \cdot \text{educat}_{it} + 0.001 \cdot \text{trad\_relig}_{it}$$

where standard errors (clustered by village) are 0.146 for the constant, 0.043 for any_rain, 0.026 for educat, and 0.104 for trad_relig. Again, none of the coefficients are statistically significant, with all having p-values greater than 0.05. The $R^2 = 0.00738$ remains low.

(iii) Both regressions suggest no statistically significant relationship between extreme weather and murders. In the first regression, the point estimate suggests that extreme rainfall is associated with 0.048 additional murders per village-year, but this effect is not statistically different from zero. Adding controls for education and traditional religion in the second regression slightly reduces the coefficient on rainfall to 0.040, and it remains insignificant. Education shows a negative relationship with murders (as expected), while traditional religion shows essentially no relationship. The very low R^2 values indicate that these variables explain almost none of the variation in murders, suggesting that other factors are more important determinants of violence in these villages.

d) (i) In the proposed IV approach, extreme rainfall (`any_rain`) serves as an instrument for famine. The first-stage regression is

$$\text{famine}_{it} = \pi_0 + \pi_1 \cdot \text{any\_rain}_{it} + v_{it}$$

which estimates the effect of extreme rainfall on the likelihood of famine. The second-stage regression is

$$\text{total\_murders}_{it} = \beta_0 + \beta_1 \cdot \widehat{\text{famine}}_{it} + \epsilon_{it}$$

which estimates the effect of famine (predicted from the first stage) on total murders. The reduced form regression is

$$\text{total\_murders}_{it} = \gamma_0 + \gamma_1 \cdot \text{any\_rain}_{it} + u_{it}$$

which directly estimates the effect of extreme rainfall on total murders.

(ii) For this IV approach to be valid, three conditions must hold:

Relevance: The instrument needs to be related strongly enough to the endogenous explanatory variable of interest. In this case, extreme rainfall should have a significant effect on the likelihood of famine. This seems appropriate, as extreme weather events can disrupt agricultural production and food availability. However, this might fail if there are other factors (e.g., government aid, market access) that mitigate the impact of rainfall, avoiding famine.

Exclusion Restriction: The instrument must not have a direct effect on the dependent variable except through its relationship with the endogenous explanatory variable. In this case, extreme rainfall should affect murders only through its impact on famine. This assumption could be appropriate if we believe that rainfall does not directly influence violence, but only through its impact on famine. However, this might fail if extreme weather also causes other stressors (e.g., displacement, resource conflicts) that directly increase violence, violating the exclusion restriction.

Exogeneity: The instrument must be uncorrelated with any unobserved variables (in the error term) that also affect the dependent variable. In this case, extreme rainfall should not be correlated with other unobserved factors that influence murders. This assumption might be appropriate if weather patterns are random and not influenced by local social or economic conditions. However, this might fail if certain villages are more prone to both extreme weather and violence due to unobserved characteristics (e.g., geographic features, historical conflicts), leading to correlation between the instrument and the error term.

Problem 2: The Primary School Deworming Project (PSDP)

For this assignment, we will analyze the dataset used in the paper titled "Worms at Work: Long-Run Impacts of a Child Health Investment."

Please download the dataset "pset3-2025-deworming.csv" from bCourses.

In this question, you will estimate the treatment effects of deworming for the following dependent variables:

1. Total years enrolled in school, 1998–2007 (`totyrs_enrolled`)
2. Indicator for passed secondary school entrance exam (`passed_primary_exam`)
3. Number of meals eaten yesterday (`num_meals_yesterday`)
4. Total hours worked in wages/self-employment/agriculture, last 7 days (`total_hours`)
5. Wages for total cash salary/food in kind, last month (`ln_emp_salary_total`)

The treatment variable varies at the school level and is called `treatment`.

The authors also include the following control variables:

```
saturation_dm+demeaned_popT_6k+zoneidI2+zoneidI3+zoneidI4+zoneidI5+zoneidI6+z
oneidI7+zoneidI8+pup_pop+month_interviewI2+month_interviewI3+month_inte
rviewI4+month_interviewI5+month_interviewI6+month_interviewI7+month_interview
I8+month_interviewI9+month_interviewI10+month_interviewI11+month_interviewI12
+cost_sharing+std98_base_I2+std98_base_I3+std98_base_I4+std98_base_I5+std98_b
ase_I6+female_baseline+avgtest96
```

Use sampling weights `weight` in your regressions.

Cluster standard errors at the school level using the variable `psdpsch98`.

a) In this question we use linear regression to estimate the effect of the deworming treatment on the five dependent variables mentioned above.

   (i) Use R to estimate regressions in the format of the following. Simply use `summary` to report the results in this question. **To receive full credits, please highlight the names of dependent variables and estimated coefficients of `treatment` with red rectangles.** You could do this by annotating the pdf document compiled from .rmd or exported from Microsoft Word.

   ```
   name_dep_var = treatment + name_control_vars, with "weight"
   as the sampling weight and "psdpsch98" as the cluster ID
   ```

   (ii) Interpret the treatment effect coefficient for the regressions on total years enrolled and passing the secondary school exam.

b) Deworming benefits might be stronger for certain groups — for instance, girls (perhaps because they were more likely to be infected) or children with lower BMI at baseline (because they were less healthy initially).

   (i) Please estimate whether the deworming treatment had a differential impact on `totyrs_enrolled`, `passed_primary_exam`, and `total_hours` by gender (`female_baseline`) and then by BMI (`BMI`). **To receive full credits, please highlight the names of dependent variables and estimated coefficients of interactive terms with red rectangles.**

   (ii) Indicate which regressions have a significant interaction term at the 10% level. Interpret the coefficient on the interaction term for these regressions.

Problem 2 continued on next page...

Solution

a) (i) Five weighted regressions with clustered standard errors were estimated. Each regression takes the form:

$$Y_i = \beta_0 + \beta_1 \cdot \text{treatment}_i + \mathbf{X}_i^\top \boldsymbol{\gamma} + \epsilon_i$$

where Y_i is the dependent variable, treatment_i is the deworming treatment indicator, and $\mathbf{X}_i$ includes all control variables. Standard errors are clustered at the school level (psdpsch98), and observations are weighted using the sampling weight. Below is a summary table showing the estimated treatment effects for each dependent variable.

	Dependent variable:				
	(1)	(2)	(3)	(4)	(5)
treatment	0.293**	0.051	0.095***	1.599	0.265***
	(0.145)	(0.031)	(0.029)	(1.036)	(0.085)
Observations	5,037	4,974	5,083	5,084	710
R^2	0.293	0.070	0.034	0.059	0.189
Note:					*p<0.1; **p<0.05; ***p<0.01

The full regression output with all control variables is included in the attached RMarkdown output. The dependent variables and treatment coefficients are highlighted with red rectangles.

(ii) The treatment coefficient on totyrs_enrolled is 0.293 with a p-value of 0.043 (statistically significant at the 5% level). This means that students in schools that received the deworming treatment were enrolled in school for approximately 0.29 additional years compared to students in control schools, holding all other factors constant. This represents an increase in educational attainment due to the deworming program.

The treatment coefficient on passed_primary_exam is 0.051 with a p-value of 0.101 (not statistically significant at the 10% level). While the point estimate suggests that treated students were 5.1 percentage points more likely to pass the primary school entrance exam, this effect is not statistically distinguishable from zero. We cannot conclude with confidence that the deworming treatment had an effect on exam passage rates.

b) (i) Six interaction regressions were estimated to test for differential treatment effects by gender and BMI. Each regression takes the form:

$$Y_i = \beta_0 + \beta_1 \cdot \text{treatment}_i + \beta_2 \cdot Z_i + \beta_3 \cdot (\text{treatment}_i \times Z_i) + \mathbf{X}_i^\top \boldsymbol{\gamma} + \epsilon_i$$

where Z_i is either female_baseline or BMI, and β_3 is the interaction coefficient of interest. Below is a summary table showing the estimated treatment effects for each dependent variable.

	Dependent variable:		
	(1)	(2)	(3)
Interaction with Gender:			
treatment:female_baseline	−0.064	0.001	−3.980**
	(0.218)	(0.040)	(2.007)
Observations	5,037	4,974	5,084
R^2	0.293	0.070	0.061
Interaction with BMI:			
treatment:BMI	−0.006	−0.001	0.075*
	(0.005)	(0.001)	(0.039)
Observations	5,017	4,955	5,064
R^2	0.294	0.071	0.060

Note: *p<0.1; **p<0.05; ***p<0.01

The full regression output with all control variables is included in the attached RMarkdown output. The dependent variables and interaction coefficients are highlighted with red rectangles.

(ii) Two regressions have significant or marginally significant interaction terms, 1) total hours worked by gender and 2) total hours worked by BMI. For total hours worked by gender, the interaction coefficient is -3.98 with a p-value of 0.047 (significant at the 5% level). This could mean that, for males (female_baseline = 0), the deworming treatment increased hours worked by 3.51 hours per week ($p = 0.018$, statistically significant). However, for females, the differential effect is -3.98 hours, making the total effect for females approximately $3.51 - 3.98 = -0.47$ hours. This indicates that the positive labor supply effect of deworming was concentrated among males, while females experienced no significant change in hours worked. This differential effect could reflect gender differences in labor market participation, household responsibilities, or the types of work opportunities available to men versus women in this context.

For total hours worked by BMI, the interaction coefficient is 0.075 with a p-value of 0.054 (marginally significant at the 10% level). This positive interaction term suggests that children with higher BMI at baseline experienced larger increases in hours worked from the deworming treatment. Specifically, for each one-unit increase in baseline BMI, the treatment effect on hours worked increased by approximately 0.075 hours per week. At the mean BMI level, the treatment effect would be close to zero (given the main effect of -0.088), but for children with BMI one standard deviation above the mean, the positive interaction effect would result in a positive net treatment effect. This finding is somewhat counterintuitive since we might expect children with lower BMI (who were less healthy) to benefit more. However, it could suggest that healthier children were better positioned to translate improved health from deworming into increased labor supply, perhaps because they had more energy reserves or faced fewer other health constraints that would limit their ability to work more hours.

None of the other interaction terms were statistically significant at the 10% level, indicating that the benefits of deworming were similar across these other demographic groups.