BST 760: Advanced Regression
Breheny

Assignment 1
Due: Thursday, January 24

1. In the 1850s, the germ theory of disease was only one of several ideas (others included "miasmas" and imbalances in the bodily humors). At the time, company-owned pumps were used as a source of drinking water. John Snow, a physician in London, attempted to show that cholera was a waterborne infectious disease by examining the connection between death rates from cholera and source of drinking water. In particular, he examined two companies, the Lambeth water company and the Southwark and Vauxhall company, that seemed alike in nearly every way (both serving the same regions of town, supplying "both rich and poor, both large houses and small," with "no difference either in the condition or occupation of the persons receiving the water" (Snow's words)), except for the fact that, in 1852, the Lambeth water company moved its intake pipe upstream in the Thames river to get purer water, while the Southwark and Vauxhall company left theirs where it was. Here are the data for the two water companies from the London cholera epidemic of 1854:

	Number of houses	Cholera deaths	Rate per 10,000
Southwark & Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37

   Snow claimed that this was convincing evidence that cholera was caused by contaminated water. His analysis, however, did not adjust for any confounders. Do you think that confounding is likely to be a problem in this study? What would be better, Snow's approach of specifically targeting these two water companies, or a study which looked at all the water companies in London, using statistical models to adjust for confounders?

2. In 1899, the early statistician Udny Yule was studying poverty. At the time, paupers in England were supported either inside grim institutions called "poor houses" or outside ("out-relief"), depending on local policy. To study the effect of these policies on poverty, Yule proposed a multiple regression equation in which the outcome variable was the change in the percentage of paupers over time, the explanatory variable of interest was the percent change in the ratio of paupers supported outside the poor house to those supported in poor houses, and explanatory variables relating to the size of the population and the percentage of elderly residents were added to control for potential confounding. Administrative districts called "unions" comprised the observations. Yule found that the coefficient pertaining to the out-relief ratio was 0.755 (*i.e.*, that a 1% increase in the out-relief ratio leads to a 0.75% increase in pauperism). Yule concluded that out-relief causes poverty.

   (a) Would you say that Yule has established causation, or merely association?

   (b) Has Yule convincingly adjusted for possible confounders, or is it possible that a confounding factor might still be the cause of the positive coefficient?

(c) What would Yule's conclusion have been if the out-relief coefficient was -0.91?

(d) What would Yule's conclusion have been if the out-relief coefficient was 0.006?

3. Extend the `ols` function that we wrote in class to return standard errors and p-values in addition to the coefficients themselves. The function should return a matrix or data frame with three columns (one for the coefficients, one for the standard errors, and one for the p-values) and a row for each coefficient. You must calculate all quantities "from scratch", in the sense that you cannot just call `lm` from within your function and return its calculations. However, by all means, test your function against the results that `lm` and `summary(fit)` return to make sure that your function is working correctly.