Ecn 102 - Analysis of Economic Data
University of California - Davis
John Parman

February 24, 2011

Problem Set 5 - Solutions

This problem set will be due by 5pm on Friday, March 11th in your TA's mailbox in the economics department mailroom. You may work in groups but everyone in the group must write up their own solutions including creating their own graphs and tables. Include any relevant regression results, calculations and graphs from Excel with your solutions but do not include the raw data.

Analyzing Energy Consumption

In this problem set, you will use multivariate regressions to analyze energy consumption by American households. The data for this problem set are contained in the file *energy-use.csv* in the data folder on Smartsite. These data are a subset of the Residential Energy Consumption Survey available through the www.data.gov website. The file on Smartsite contains data on single-family detached houses and has the following set of variables:

- **hd65** - heating degree-days (this is a measure of how much heating is required over the year to warm the house up to 65 degrees, a decrease in the outside temperature of one degree for one day of the year would increase hd65 by one unit)
- **cd65** - cooling degree-days (this is a measure of how much cooling is requried over the year to cool the house down to 65 degrees, an increase in the outside temperature of one degree for one day of the year would increase cd65 by one unit)
- **totrooms** - number of rooms in the house
- **yearbuilt** - year in which the house was built
- **washload** - number of loads of laundry done each week (this is not the exact definition of this variable in the survey but it will work for our purposes)
- **kwh** - kilowatt-hours of electricity used annually
- **solar** - solar power dummy (equals one if household uses solar power for any purpose, equals zero if household does not use solar power)

a. Most households use air conditioning powered by electricity to cool down the house but use other forms of energy (gas, oil, etc.) to warm up the house. Given this piece of information, what would you predict for the sign and significance of the coefficients if electicity usage were regressed on a household's heating requirements and cooling requirements? Run a regression of electricity usage (kwh) on heating degree-days (hd65) and cooling degree-days (cd65). Are your results consistent with your predictions?

For cooling degree-days, we would expect a positive sign and most likely a highly significant coefficient. More cooling degree-days would mean more air conditioner use which would lead to higher electricity usage. For heating degree-days, we would also likely expect a positive coefficient (some households use electricity for heat so greater heating requirements would lead to greater electricity usage). However, the coefficient might not be as significant if most houses are heated by means other than electricity. The regressions (see ps5-solution.xlsx) confirm these guesses. The coefficient for cd65 is positive and highly statistically significant. The coefficient for hd65 is positive but not statistically significant at any reasonable significance level.

b. We'll focus on the relationship between cooling degree-days and electricity usage for the remainder of the problem. We could just run a simple regression of electricity usage (kwh) on cooling degree-days (cd65). However, we have additional variables that we could include that will affect electricity usage and would improve the explanatory power of our regression. Run a regression of electricty usage (kwh) on cooling degree-days (cd65), washer loads per week (washload), rooms in the house (totrooms) and whether the house uses solar power (solar). Are all of the signs of the coefficients what you would expect? Explain your answer for each coefficient.

The signs of the coefficients are all quite intuitive. We've already discussed cd65 above. For washer loads, we get a positive coefficients. Washing machines require electricity, so doing one more load of laundry a week should increase electricity usage giving us the observed positive coefficient on washload. Larger houses will require more energy to heat, cool, light, etc. So it makes sense that we get a positive coefficient for totrooms. Finally, houses with solar panels will be generating their own electricity and therefore won't need to draw as much electricity from the grid, giving us the negative coefficient on solar.

c. An average washing machine uses a little less than one kilowatt-hour of electricity per load. Assuming that one load of laundry requires one kilowatt-hour of electricity, what would you expect the magnitude of the coefficient on washer loads per week to be? How does this compare to the coefficient you got for the number of washer loads? Give a possible explanation for any discrepencies between the size of your actual coefficient and the magnitude you expected based on the electricity use of an average washing machine. (Hint: Think about other variables that affect electricity usage and are not included in our regression but are correlated with the number of loads of laundry).

If I increase washer loads per week by one, that would increase electricity usage by one kilowatt-hour per week or 52 kilowatt-hours per year. So if the only difference in electricity usage associated with an extra washer load is the electricity actually used by the washing machine, we would expect to

get a coefficient on washload of 52. Our estimated coefficient is much larger than this. An explanation is that we have an upward bias coming from some omitted variable. An obvious omitted variable will be family size. Larger families will do more laundry. However, they will also use more electricity for all sorts of other reasons. So family size is an omitted variable that is positively correlated with washload and positively correlated with kwh. This will produce a positive bias for the coefficient on washload.

d. Would you expect the amount of electricity generated annually by a house's solar panels to be less than, equal to or greater than the value of the estimated coefficient for the solar dummy? Explain your answer.

If solar panels were randomly assigned to people, then we would probably expect the coefficient on solar panel to be exactly equal in magnitude to the amount of electricity generated by the solar panels. For each extra kilowatt-hour of electricity generated by the solar panels, electricity drawn from the power grid would decrease by one kilowatt-hour. However, solar panels aren't randomly assigned. There will be unobserved charactersitics of households that are correlated with having solar panels and will lead to omitted variable bias. For example, suppose the people that install solar panels are the people that are more environmentally conscious. These people may be more inclined to conserve energy in all aspects of their living. So we have an omitted variable, let's call it 'environment consciousness', that is positively correlated with the solar dummy variable and negatively correlated with the outcome variable kwh. So there will be a negative bias on the solar coefficient. So the coefficient will likely be larger in magnitude than the amount of power generated by the solar panels.

e. It takes different amounts of energy to cool different houses down by one degree. For example, a large house will require more energy to cool down by a degree because a greater volume of air needs to be cooled. We can capture these effects through interaction terms. Create an interaction term between the number of rooms and cooling degree-days and another interaction term between the number of washer loads and the cooling degree-days. Rerurn your regression from part (b) including these interaction terms. Are the signs and significance of the coefficients on the interaction terms what you would expect? Why or why not? (You can assume that washer loads are a proxy for family size. In other words, the important change captured by an increase in washer loads is an increase in family size.)

See ps5-solutions.xlsx for the regressions. The sign for the cd65-totrooms coefficient is positive and the coefficient is statistically significant. This should make sense. If the temperature increases by one degree, it is going to take more electricity to cool down a large house that one degree than it will to

cool down a small house. The coefficient for the cd65-washload interaction term is also positive, suggesting that an increase in temperature leads to a greater increase in electricity usage for a household with many people than a household with just a few people. This could make sense if we think of more people requiring more fans and more air conditioners when temperatures rise. However, the coefficient is not statistically significant. This isn't all the surprising as the case for strong interaction isn't nearly as strong as it was for the number of rooms and cooling degree-days.

f. Use an F test to determine whether including these interaction terms improved the regression. In other words, test whether the coefficients for the interaction terms are both zero or whether at least one of the coefficients is different from zero.

See ps5-solutions.xlsx for the F-test. The p-value is incredibly small so we will reject the null hypothesis that both interaction term coefficients are zero at any reasonable significance level.

g. Based on your regression results from part (e), draw a line graph showing the predicted value of annual electricity usage as a function of the number of cooling degree-days for a household with five rooms that does one load of laundry a week and does not use solar power. Label the slope and intercept of your line with the appropriate values. On the same graph, draw another line showing the predicted value of annual electricity usage as a function of the number of cooling degree-days for a household with ten rooms that does one load of laundry a week and does not use solar power. Label the slope and intercept. You can draw this graph by hand. (It is not important that your graph is drawn to scale. What is important is that you label the slopes and intercepts correctly.)

Our regression equation that we estimated in Excel is:

$$kwh = b_1 + b_2 solar + b_3 cd65 + b_4 totrooms + b_5 washload +$$
$$b_6 cd65 \cdot totrooms + b_7 cd65 \cdot washload$$

Plugging in the values of the coefficients from our regression results, this equation becomes:

$$kwh = 267 - 3456 \cdot solar - .94 \cdot cd65 + 573 \cdot totrooms + 1883 \cdot washload +$$
$$.53 \cdot cd65 \cdot totrooms + .10 \cdot cd65 \cdot washload$$

Plugging in 0 for $solar$, 5 for $totrooms$ and 1 for $washload$ gives us:

$$kwh = 267 - 0 - .94 \cdot cd65 + 573 \cdot 5 + 1883 \cdot 1 +$$
$$.53 \cdot 5 \cdot cd65 + .10 \cdot 1 \cdot cd65$$

Simplifying this expression leaves us with:

$$kwh = 5015 + 1.81 \cdot cd65$$

If we use 10 for *totrooms* instead of 5, we would get the following:

$$kwh = 7880 + 4.46 \cdot cd65$$

Graphing these two equations gives us: