

In this session we will work through examples using single variable regression.

AIR QUALITY & ASTHMA

We begin with the asthma.xls data on the website. Input these data into JMP. The data provide information on air quality and other characteristics for 50 U.S. cities. The following table defines the variables.

<u>Variable</u>	<u>Definition</u>
city	The name of the city
badairdays	The percentage of summer days with air quality that is dangerous for children with asthma in 2002
lnbadairdays	The natural log of badairdays
chbadairdays	The change in badairdays from 1998 to 2002
asthma_attacks1	The percentage of children in the city reported to have chronic asthma attacks in 2002
asthma_attacks2	The same as asthma_attacks1 except for the value for Riverside has been mistakenly entered as '2'
asthma_attacks3	The same as asthma_attacks1 except that data for the 5 cities with the worst air quality are dropped
lnasthma_attacks1	The natural log of asthma_attacks1
home_price	The median price of a home in 2002 (in \$000s)
chhome_price	The change in median home price from 1998 to 2002

- 1 We begin by exploring the relationship between asthma attacks and air quality. Draw a scatter-plot with asthma_attacks1 on the Y axis and bad air days on the X axis.
 - a) Visually, does the covariance of X and Y appear to be negative or positive?
 - b) Run a regression of $Y = \text{asthma_attacks1}$ on $X = \text{badairdays}$. Interpret the following:
 - i. R^2
 - ii. The slope coefficient
 - iii. The intercept
 - iv. Evaluate the null hypothesis that asthma attacks are unrelated to outdoor air quality

ESM 206 Lab Exercise #3
Week 9, Winter Quarter

- c) Run a regression of $Y = \text{asthma_attacks2}$ on $X = \text{badairdays}$ and compare your interpretations with i-iv above. What are the effects of making this single data-entry mistake (and being correct in the other 49 instances)?

- d) Redo the $Y = \text{asthma_attacks3}$ on $X = \text{badairdays}$ and compare your interpretations with i-iv above. Notice that now we have dropped our observations from the 5 cities with the worst air quality. How do the results compare with (b)? What do the results from (b) and (d) suggest about the ‘localness’ of the linear relationship between air quality and asthma attacks?

- e) Because there is some evidence that the relationship between X and Y is nonlinear (at least over the full range of the data) we log Y and X .

Run a regression of $Y = \ln \text{asthma_attacks1}$ on $X = \ln \text{badairdays}$. Interpret the following:

- i. R^2
- ii. The slope coefficient
- iii. Evaluate the null hypothesis that asthma attacks are unrelated to outdoor air quality

2 Economic theory suggests that air quality in a locale will be capitalized into land values and home prices. In the questions provide some exploratory analysis of this theory.

- a) Run a regression where $Y = \text{home_price}$ and $X = \text{badairdays}$. Interpret the following:
 - i. R^2
 - ii. The slope coefficient
 - iii. The intercept
 - iv. Evaluate the null hypothesis that asthma attacks are unrelated to outdoor air quality

- b) Do the results from (a) convince you that there is no relationship between air quality and home price? What are some of the shortcomings of this analysis in terms of inferring causation?

- c) Suppose that in 1998 the federal EPA deemed some of the cities out of compliance with the Clean Air Act. Suppose also that the EPA forced the non-complying cities to reduce particulate air matter from 1998-2002; in some cases the air actually became cleaner in the non-complying cities by 2002.

Run a regression where $Y = \text{chhome_price}$ and $X = \text{chbadairdays}$. Interpret the following:

- i. R^2
 - ii. The slope coefficient
 - iii. The intercept
 - iv. Evaluate the null hypothesis that asthma attacks are unrelated to outdoor air quality
- d) Do the results from (c) suggest there is in fact a relationship between air quality and home price? How does this analysis overcome some of the shortcomings from (a) in terms of assessing causation?
- e) What are some of the shortcomings of the analysis in (c) in terms of inferring causation?