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

We begin by examining a slightly modified version of the asthma data that we analyzed in the last lab session. For this session use the asthma2.xls file on the website.

Variable	Definition
city	The name of the city
badairdays	The percentage of summer days with air quality that is dangerous for children with asthma in 2002
asthma_attacks	The percentage of children in the city reported to have chronic asthma attacks in 2002
per_smokers	The percent of adults in the city who smoke
per_adults	The percent of the adults in the city reported to have asthma symptoms

1 As before our goal is to test the null hypothesis that outdoor air quality has no effect on asthma in children. Unlike before we now have data on the percentage of adults who smoke in each city. We also have data on the percentage of adults with asthma symptoms, and we want to control for this because asthma may be hereditary.

a) Run a regression where Y= asthma_attacks and control for all of the *X* variables. Can we reject the null hypothesis that outdoor air quality has no effect on asthma in children?

[Analyze->Fit Model->Y is asthma_attacks, Add badairdays and per_smokers and per_adults -> Run Model]

We fail to reject the null hypothesis that outdoor air quality has no effect on asthma incidence in children.

b) Look at the residuals by predicted plot. Is there visual evidence that the assumptions of OLS are violated?

No, the errors look well-behaved.

c) Why may it be problematic to include the per_adults variable along with badairdays in your regression? Examine the correlation among your predictor variables. Is there evidence that per_adults and badairdays are highly correlated?

[Analyze ->Multivariate Methods-> Multivariate -> Y is badairdays and per_smokers and per_adults]

Yes, the correlation between per_adults and badairdays is 0.90

d) Run the regression without per_adults. Now are you able to reject the null hypothesis that outdoor air quality has no effect on asthma in children? If your answer differs from (a), explain why.

[Analyze->Fit Model->Y is asthma_attacks, Add badairdays and per_smokers and per_adults -> Run Model]

Now we soundly reject the null hypothesis that air quality has no effect on asthma attacks. We have a different result than (b) because of the high collinearity between badairdays and per_adults.

For the next question we'll use the cars_1993.xls data on the website. These are the data used in Homework 3 and the table below describes the variables.

Variable	Definition
Manufacturer	Manufacturer name
Model	Car model
Type	Car type
Price	Price in $000s
CityMPG	MPG in the city
HighwayMPG	MPG on highways
EngineSize	In liters
Horsepower	Horsepower
Fuel Tank	Capacity of tank in gallons
Passengers	Number of passengers car fits
Weight	In pounds
Origin	Whether or not the car is an American brand

2 We begin by examining how City MPG differs between foreign and US cars. First, we want to convert the 'Origin' categorical variable into a numerical variable. Create a new column by right-clicking on the top right of the data sheet. Click Column Properties->Formula->Conditional->If. The conditional expression should read: If Origin= = "US" then 1, else 0. Name this variable "Origin_Dummy" and leave it as a continuous variable.

a) Record the means for CityMPG for US versus foreign cars. Which is higher?

[Analyze-> Distribution ->CityMPG by Origin_Dummy]

Origin	*Mean of CityMPG*
Foreign	*23.98*
US	*20.96*

b) Now run a regression of Y=CityMPG on X= Origin_Dummy. Interpret the results and show how all of the information in (a) is conveyed in this regression. Can you reject the null hypothesis that the mean CityMPG for US cars is no different than the mean for foreign cars?

[Analyze-> Fit Y by X ->Fit Line]

The mean CityMPG is 23.98 if X = 0.
The mean CityMPG is 23.98-3.019 = 20.96 if X=1.

We can reject the null hypothesis that the means do not differ at $P \leq 0.01$.

c) Recall that in Homework 3 we examined separately the relationship between weight and fuel efficiency by origin (e.g., we ran regressions of Weight on CityMPG for US and foreign cars). Now we can more rigorously test for whether the effect of Weight differs by origin by including an interaction (Weight*Origin_Dummy) in our regression. Our updated model is

$$\text{CityMPG} = \beta_0 + \beta_1(\text{Origin\_Dummy}) + \beta_2(\text{Weight}) + \beta_3(\text{Origin\_Dummy*Weight}) + \varepsilon$$

Which coefficient signifies the effect of weight on CityMPG for foreign cars? For US cars? Does the regression suggest that the fuel efficiency of US or foreign cars are more sensitive to weight?

*[Analyze->Fit Model->Y is CityMPG and Add Origin_Dummy and Weight and Origin_Dummy*Weight -> Run Model]*

The predicted effects are given in the table below

	US Cars	*Foreign Cars*
Origin	*1*	*0*
CityMPG	$\beta_0 + \beta_1 + \beta_2 * Weight + \beta_3 * Weight =$ $\beta_0 + \beta_1 + (\beta_2 + \beta_3) * Weight$	$\beta_0 + \beta_2 * Weight$
Effect of weight	$\beta_2 + \beta_3$	β_2
Marginal effect of weight on US cars	β_3	---

Because β_3 is negative and statistically different than zero, we find that the fuel efficiency of US cars is more sensitive to weight.