Regression & ANCOVA

In this exercise, we will learn how to perform linear regression using software. We will use the dataset `Lung.xls` posted on Blackboard. This dataset describes the Childhood Respiratory Health survey discussed in Ch 15 notes. Measures collected by this survey include FEV_1 (Liters/second), age, school, gender and smoking status.

In this analysis, we would like to investigate the relationship between FEV_1 and age - how does age affect FEV_1 among children? How does this relationship change when we consider smoking status?

The SAS code in this exercise can be found in the file Lab11.sas, which is posted on Blackboard.

1 SAS

1.1 Read In The Data

We will read in the dataset `Lung.xls` using the SAS Import Wizard. Follow the steps below to import the dataset into SAS.

1. Make sure `Lung.xls` file is closed outside of SAS, then open up SAS.
2. From the File menu in SAS, select 'Import Data ...'.
3. Check 'Standard Data Source' and select 'Microsoft Excel 97/2000/2002/2003 Workbook' from the pulldown menu.
4. When the window pops up, click Browse and locate your file ('H:\bios600\Data\Lung.xls'). Click Open and then OK.
5. Select the table 'fev$' (Sheet title) from the pulldown menu and hit Next.
6. Select 'WORK' to put the file in the Work library. Name your SAS dataset in the 'Member:' box by typing 'Lung'.
7. Click Finish.

After completing these steps, click on the Log. You should see the following statement in blue.

```
NOTE: WORK.LUNG data set was successfully created.
```

1.2 Descriptive Statistics

Our first step is to explore the data using descriptive statistics on both variables `fev1` and `age`. We are interested in describing these variables separately for each smoking status. We use the `MEANS` procedure for this.

```
*fev1 and age stats overall ;
PROC MEANS DATA=pulmonary N NMISS MEAN STD MIN MAX MAXDEC=4;
  VAR fev1 age;
RUN ;

*fev1 and age stats separately by smoke ;
PROC MEANS DATA=pulmonary N NMISS MEAN STD MIN MAX MAXDEC=4;
  CLASS smoke ;
  VAR fev1 age;
RUN ;
```

1.3 Simple Linear Regression

Suppose we are willing to assuming that age and FEV_1 are related in a linear fashion. Then we would like to fit the simple linear regression model

$$Y_i = \alpha + \beta X_i + \epsilon_i \qquad \text{(in math terms)}$$

$$\text{FEV1}_i = \alpha + \beta(\text{AGE}_i) + \epsilon_i \qquad \text{(in variable terms)}$$

where we assume $\epsilon_i \sim N(0, \sigma)$. Then we would like to assess the strength of this linear relationship between X=age and Y=FEV_1. To do this, we will test the null hypothesis of "no linear relationship", $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$ at $\alpha = 0.05$ level of significance.

We fit the regression model $\hat{y}_i = a + bx_i$ (i.e., get the fitted model) and carry out the hypothesis test by using the `REG` procedure.

```
PROC REG DATA=lung ;
  MODEL fev1 = age / CLB;
  PLOT fev1*age ;
RUN ; QUIT ;
```

Assuming everything runs correctly, SAS will fit the model and perform the desired 2-sided hypothesis test about the slope parameter β and generate the results. The output is shown on the next page.

The `MODEL` statement is written in a very specific way that mirrors the assumed model, $Y = X$. This will always be the format of a `MODEL` statement, since this is where we are actually telling SAS the specific model we want to fit.

Additionally, the `PLOT` statement immediately below is written specifically to mirror our assumed model. This statement tells SAS to also provide a scatterplot of Y=FEV_1 by X=age (response*explanatory).

The REG Procedure
Model: MODEL1
Dependent Variable: fev1 fev1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	280.91916	280.91916	872.18	<.0001
Error	652	210.00068	0.32209		
Corrected Total	653	490.91984			

Root MSE	0.56753	R-Square	0.5722
Dependent Mean	2.63678	Adj R-Sq	0.5716
Coeff Var	21.52349		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.43165	0.07790	5.54	<.0001	0.27868	0.58461
age	1	0.22204	0.00752	29.53	<.0001	0.20728	0.23680

SAS output often includes lots of things you don't need. The most important things that you'll need from this output are

1. First table - ANOVA Table
2. Second table
   - Root MSE $= \sqrt{MSE} = \sqrt{MS_W}$. This is the estimate $s_{Y|X} = \hat{\sigma}$ of σ from model assumption $\epsilon_i \sim N(0, \sigma)$.
   - R-Square Coefficient of Determination (r^2). Can be used to get $|r|$. To get r, you need to look at a plot of the data to determine the direction of the correlation.
3. Third table - Parameter Estimate Table
   - Estimates $a,\ b$ of $\alpha,\ \beta$ to write down regression equation.
   - Standard errors of estimates (SE_a, SE_b).
   - Hyp. tests of $H_0 : \alpha = 0$ and $H_0 : \beta = 0$.
   - 95% CI's for α and β.

1.3.1 Varying Significance Level

The tests in the previous section were conducted at the significance level $\alpha = 0.05$, which is the default level. We can change this in the `PROC` statement by adding the SAS Keyword option `alpha=`, along with whatever value we prefer. For instance, to use $\alpha = 0.01$ and calculate the corresponding 99% confidence intervals for α and β, we would use the following code.

```
PROC REG DATA=lung alpha=0.01;
  MODEL fev1 = age / CLB;
  PLOT fev1*age ;
RUN ; QUIT ;
```

1.4 Multiple Linear Regression (ANCOVA)

Suppose we now want to fit the same linear regression model as in §1.3, but we want to adjust for the effects of smoking status. That is, we want to understand the linear relationship between Y=FEV$_1$ and X=age in the presence of the covariate `smoke`. Now, the linear model is

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \qquad \text{(in math terms)}$$

$$\text{FEV1}_i = \alpha + \beta_1 \text{AGE}_i + \beta_2 \text{SMOKE}_i + \epsilon_i \qquad \text{(in variable terms)}$$

where X_1=age and X_2=smoke. The fitted model is now $\hat{y}_i = a + b_1 x_1 + b_2 x_2$, and we fit it in the same way as the simple linear regression model. This time, we add `smoke` to the `MODEL` statement, on the righthand side of the equation.

```
PROC REG DATA=lung ;
  MODEL fev1 = age smoke / CLB;
RUN ; QUIT ;
```

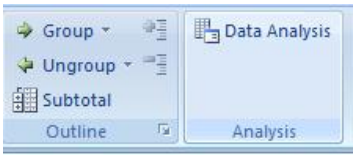
Now, any conclusions we make about the relationships of either `age` or `smoke` with FEV$_1$ must be interpreted carefully. That is, each of these relationships must be interpreted in the context (or presence) of the other variable. See Chapter 15 notes for discussion.

2 Excel

2.1 Load Analysis ToolPak

First, open up the dataset `Lung.xls` in Excel. Click on the word 'Data' at the top of the screen to open up the Data ribbon ('Data' is fifth from left, starting with the word 'Home').

With the Data ribbon open, look at the right-most section of the ribbon menu. If the last two sections are 'Outline' and 'Analysis' and look like the picture below, then the Analysis ToolPak is ready to use.



If the last section of your Data ribbon only has the 'Outline' section but NOT the 'Analysis' section, then use the following steps to load the Analysis ToolPak in Excel 2007. (Directions for other versions of Excel can be found in Week1a_GetSoftware.pdf posted on Blackboard under Course Documents → Lab Exercises → Week 1.)

1. Click the Microsoft Office Button (top left corner)  , and then click Excel Options (bottom right).
2. Click Add-Ins from menu at left, and then in the Manage box (bottom), select Excel Add-ins.
3. Click Go.
4. In the Add-Ins available box, select the Analysis ToolPak check box, and click OK.

* If Analysis ToolPak is not listed in the Add-Ins available box, click Browse to locate it.

** If you get prompted that the Analysis ToolPak is not currently installed on your computer, click Yes to install it.

The Analysis ToolPak should now be loaded. Click on the Data Ribbon again, and confirm that the last two sections at right are 'Outline' and 'Analysis', as in the picture above.

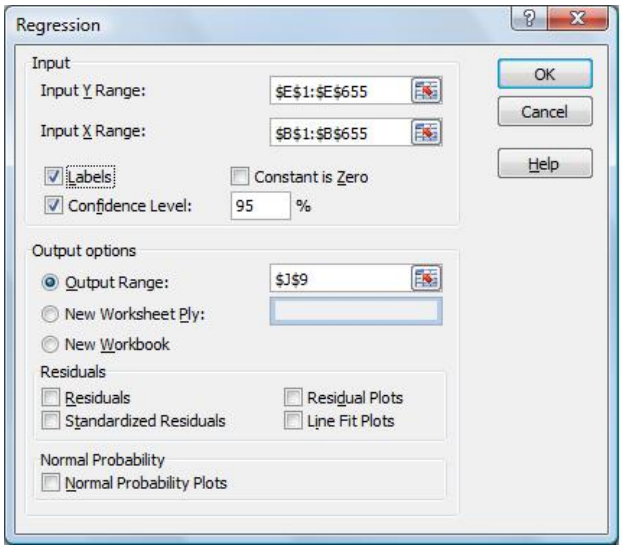
2.2 Simple Linear Regression

As before, we want to investigate the relationship between Y=fev1 and X=age with the linear model

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

where we assume $\epsilon_i \sim N(0, \sigma)$. We can use the Regression tool from the Analysis ToolPak in order to fit the model and conduct the desired t-test about the slope parameter. These are the steps we shall use to carry out this analysis. The parameters used in this exercise are shown in the picture below.

1. Open the Data Analysis dialog box from the Data ribbon.
2. Highlight "Regression" and press OK.
3. Set "Input Y Range" to be the cells containing FEV_1 data, including the label.
4. Set "Input X Range" to be the cells containing AGE data, including the label.
5. Check the box "Labels".
6. Check the box "Confidence Level" and set to 95%.
7. Decide where you want the output to print (cell J9). Press OK.



2.3 Multiple Linear Regression

Now we want to investigate the relationship between Y=fev1 and X_1=age in the presence of the variable X_2=smoke. To assess this relationship while accounting for the effect of smoking status, we will use the linear model

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

where we assume $\epsilon_i \sim N(0, \sigma)$. We can use the Regression tool from the Analysis ToolPak in order to fit the model and conduct the desired t-test about the slope parameter. These are the steps we shall use to carry out this analysis. The parameters used in this exercise are shown in the picture below.

1. Open the Data Analysis dialog box from the Data ribbon.

2. Highlight "Regression" and press OK.
3. Set Input Y Range to be the cells containing FEV_1 data, including the label.
4. Set Input X Range to be the cells containing AGE & SMOKE data, including the labels.
5. Check the box "Labels".
6. Check the box "Confidence Level" and set to 95%.
7. Decide where you want the output to print (cell J30). Press OK.