Lab 1 – Spring 2021

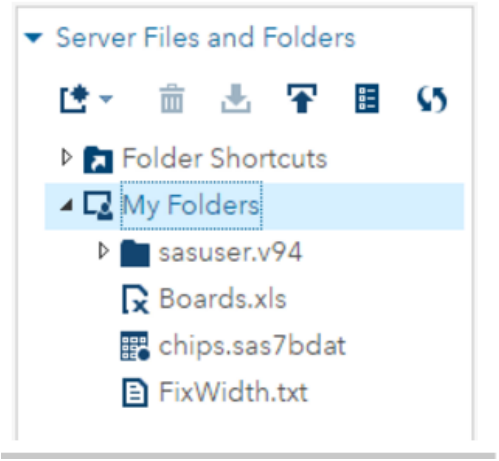
04^{th} March 2021
cavicchia@ese.eur.nl

1 Get Ready to Start

How to import data in SAS?
Before you can import the data, you must be able to access the data file from SAS Studio. If you are running SAS University Edition from virtualization software such as VirtualBox, you must save these files to a **shared folder**.

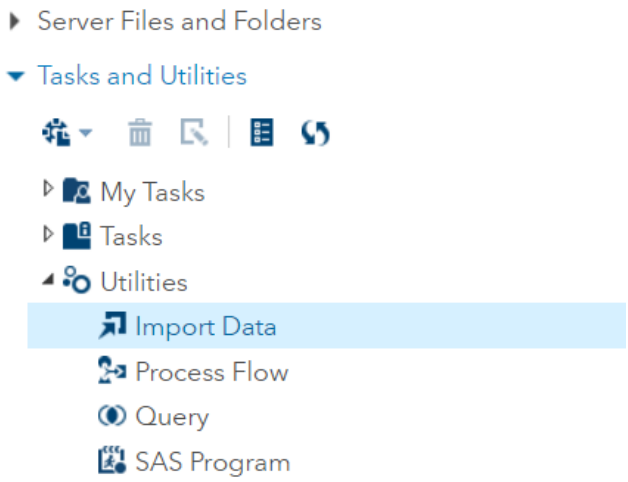
The shared folder should contain the following content:

- any preferences or settings that you specify in SAS Studio
- any data and results that you want to access from the SAS University Edition and your local computer



Before you start working in the SAS University Edition, you should create a shared folder called `myfolders`. The content in a shared folder persists between sessions and is preserved when the SAS University Edition is updated. In SAS Studio (which is the user interface for SAS University Edition), the `myfolders` shared folder is available as **Folders** > **My Folders**.
To access this shared folder from a SAS program, use `/folders/myfolders`. This is the logical location for the shared folder in the SAS University Edition. The physical location of this shared folder depends on your operating environment. For example, in Windows operating environments, this physical path could be `C:\SASUniversityEdition\ myfolders`. You can use the SAS University Edition without creating the `myfolders` shared folder. However, any content that you save might be lost when the SAS University Edition is updated. In addition, a warning message appears on the SAS University Edition Information Center.

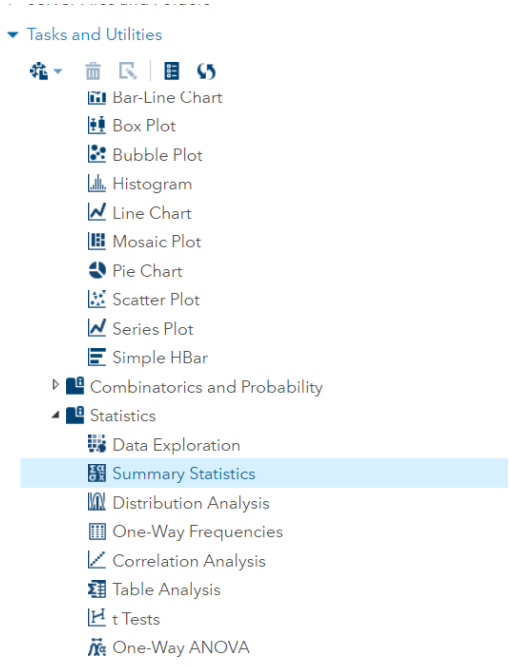
Now we are ready to import the data. Select the bottom Run and see the output data. This dataset of size $n = 51$ are for the 50 states and the District of Columbia in the United States - (Data source: Mind On Statistics, 3rd edition, Utts and Heckard). There are 6 columns (1 label column and 5 variables):



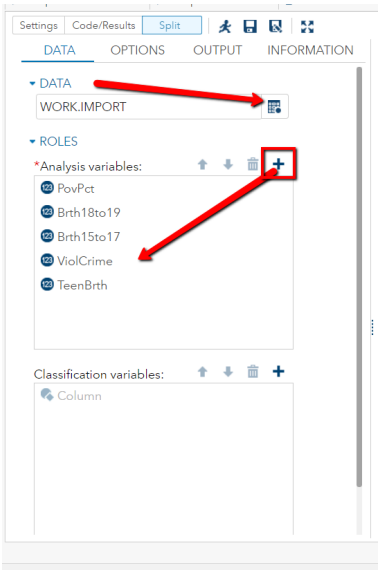
- **Location** - State name;
- **PovPct** - Percentage of population living in households with income below *poverty level*;
- **Brth15to17** - Birth rate for females 15 to 17 years old = births per 1,000 persons in group;
- **Brth18to19** - Birth rate for females 18 to 19 years old = births per 1,000 persons in group;
- **ViolCrime** - Violent crime rate in state;
- **TeenBrth** - Birth rate for females 15 to 19 years old = births per 1,000 persons in group.

2 Descriptive Statistics

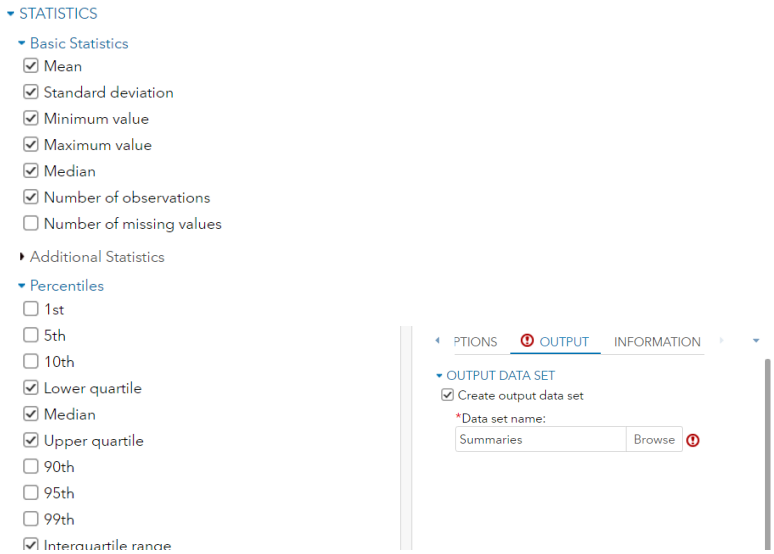
First, we examine the main descriptive statistics of the data. The Summary Statistics task provides data summarization tools to compute descriptive statistics for variables across all observations and within groups of observations. You can also summarize your data in a graphical display, such as a histogram. These can be obtained as follows. First select **Summary Statistics** under **Statistics** in the **Tasks and Utilities**. Then, the user interface for the Summary Statistics task opens.



On the Data tab, select the WORK.IMPORTDATA dataset(it is possible to rename the dataset, if you like another name!). To the Analysis variables role, assign the column names, that are the variables, as depicted in the figure



Under the window **Statistics**, select the following summary descriptives, for example. To see the output, you have to give a name for the output, as shown below (by selecting the window OUTPUT). To get the summary statistics, select the bottom run (or F3).



2.1 Output

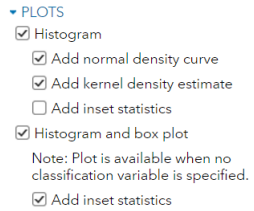
The summary statistics are the following ones

Variable	Mean	Std Dev	Minimum	Maximum	Median	N	Lower Quartile	Upper Quartile	Quartile Range
PovPct	13.1176471	4.2772283	5.3000000	25.3000000	12.2000000	51	10.2000000	16.1000000	5.9000000
Brth18to19	72.0196078	18.9755634	39.0000000	104.3000000	69.4000000	51	57.6000000	88.7000000	31.1000000
Brth15to17	22.2823529	8.0434994	8.1000000	44.8000000	20.0000000	51	17.2000000	28.2000000	11.0000000
ViolCrime	7.8549020	8.9141307	0.9000000	65.0000000	6.3000000	51	3.9000000	9.5000000	5.6000000
TeenBrth	42.2431373	12.3185105	20.0000000	69.1000000	39.5000000	51	33.0000000	53.0000000	20.0000000

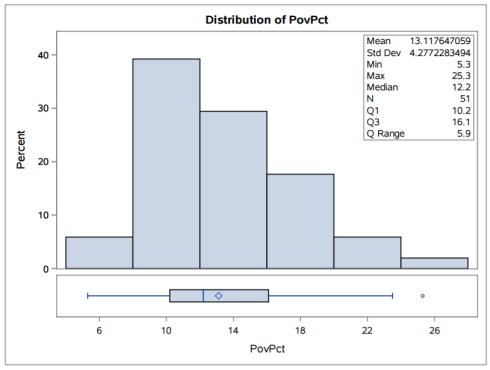
We see that Birth18to19 has highest mean, standard deviation, and IQR. Although ViolCrime shows the largest range, its IQR is the smallest. We deduce that there may be very values (indeed mean is larger than median) that affect the range, standard deviation and mean. Median and IQR are not affected by extreme values (that can be positive or negative). For this reason, they are said to be robust summary statistics. They give us an idea about the center of the distribution and its variability (and its shape - more or less concentrated around the central value), respectively. Comparing the means with the medians, we deduce that the distributions are quite symmetric for all variables, apart from ViolCrime.

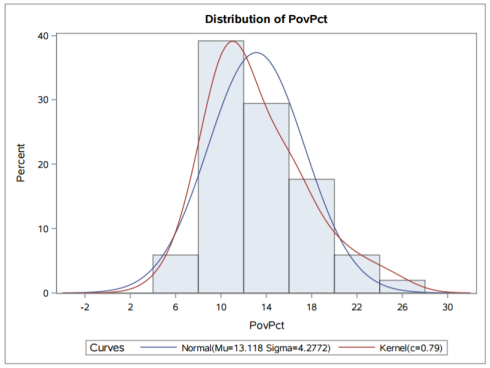
2.2 Plots

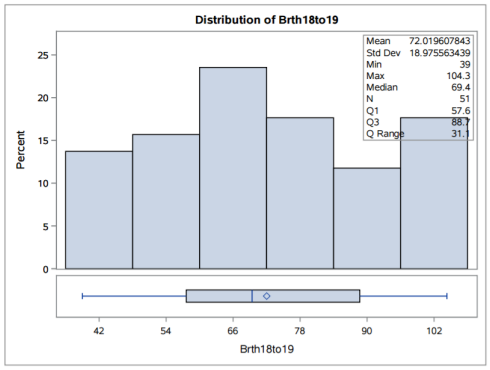
To complete the univariate analysis, it is possible to create some plots, histograms and boxplots, as follows.
Under the window **Plot**, select the following plots, for example.

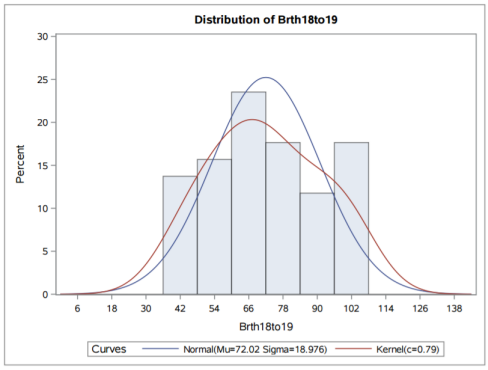


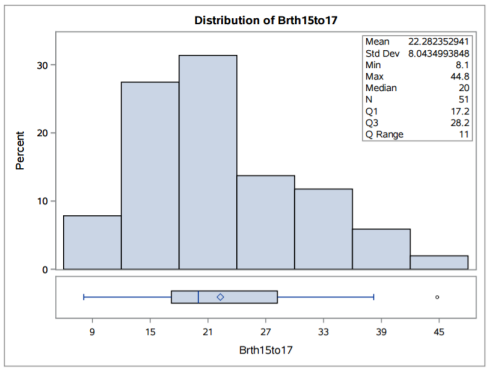
To see the output, you have to give a name for the output, as shown below (by selecting the window OUTPUT). To get the plots, select the bottom run (or F3).
The plots depict graphically the main features of the variables. The distributions of the variables PovPct and TeenBirth result to be approximately normally distributed well approximated (the corresponding normal density and kernel density are very close to each other). The distribution of the variable ViolCrime departs from the normality distribution more heavily. The boxplot summarizes the variable distribution in terms of quartiles, but it does not show some important features, such as multimodality, presence of outliers, gaps between values. On the other hand, the histogram does not put in evidence how the data are distributed in terms of quartiles. It is always advisable to look at both plots to draw univariate conclusions about the data.

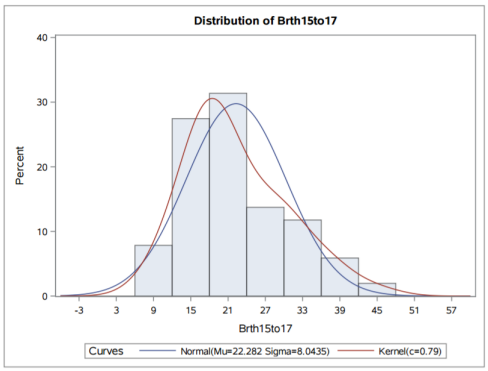

Distribution of PovPct
Mean 13.117647059
Std Dev 4.2772283494
Min 5.3
Max 25.3
Median 12.2
N 51
Q1 10.2
Q3 16.1
Q Range 5.9
Percent
PovPct

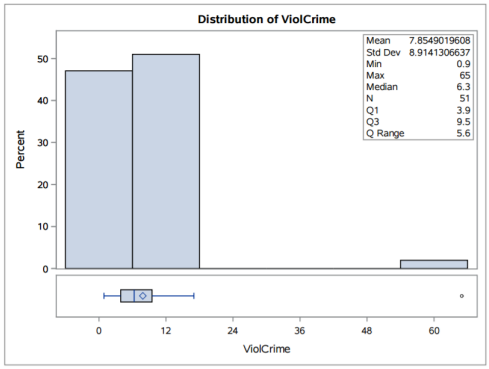

Distribution of PovPct
Percent
PovPct
Curves
Normal(Mu=13.118 Sigma=4.2772)
Kernel(c=0.79)

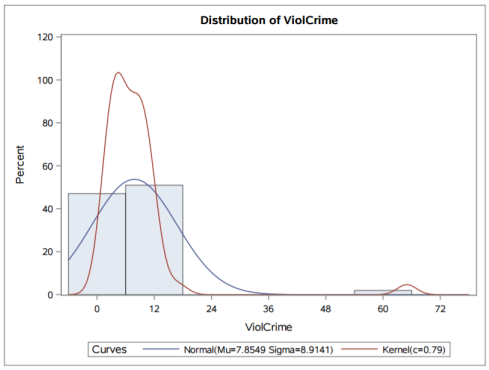

Distribution of Brth18to19
Mean 72.019607843
Std Dev 18.975563439
Min 39
Max 104.3
Median 69.4
N 51
Q1 57.6
Q3 88.7
Q Range 31.1
Percent
Brth18to19

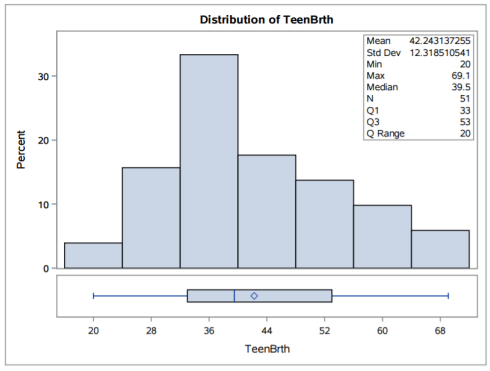

Distribution of Brth18to19
Percent
Brth18to19
Curves
Normal(Mu=72.02 Sigma=18.976)
Kernel(c=0.79)

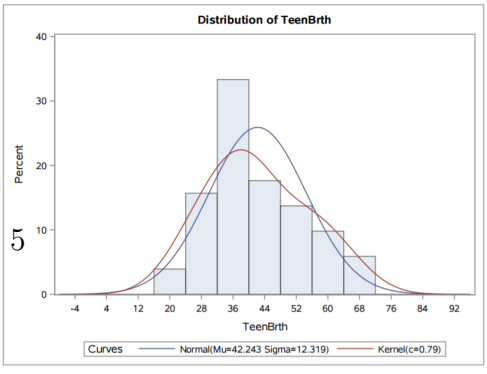

Distribution of Brth15to17
Mean 22.282352941
Std Dev 8.0434993848
Min 8.1
Max 44.8
Median 20
N 51
Q1 17.2
Q3 28.2
Q Range 11
Percent
Brth15to17


Distribution of Brth15to17
Percent
Brth15to17
Curves
Normal(Mu=22.282 Sigma=8.0435)
Kernel(c=0.79)


Distribution of ViolCrime
Mean 7.8549019608
Std Dev 8.9141306637
Min 0.9
Max 65
Median 6.3
N 51
Q1 3.9
Q3 9.5
Q Range 5.6
Percent
ViolCrime

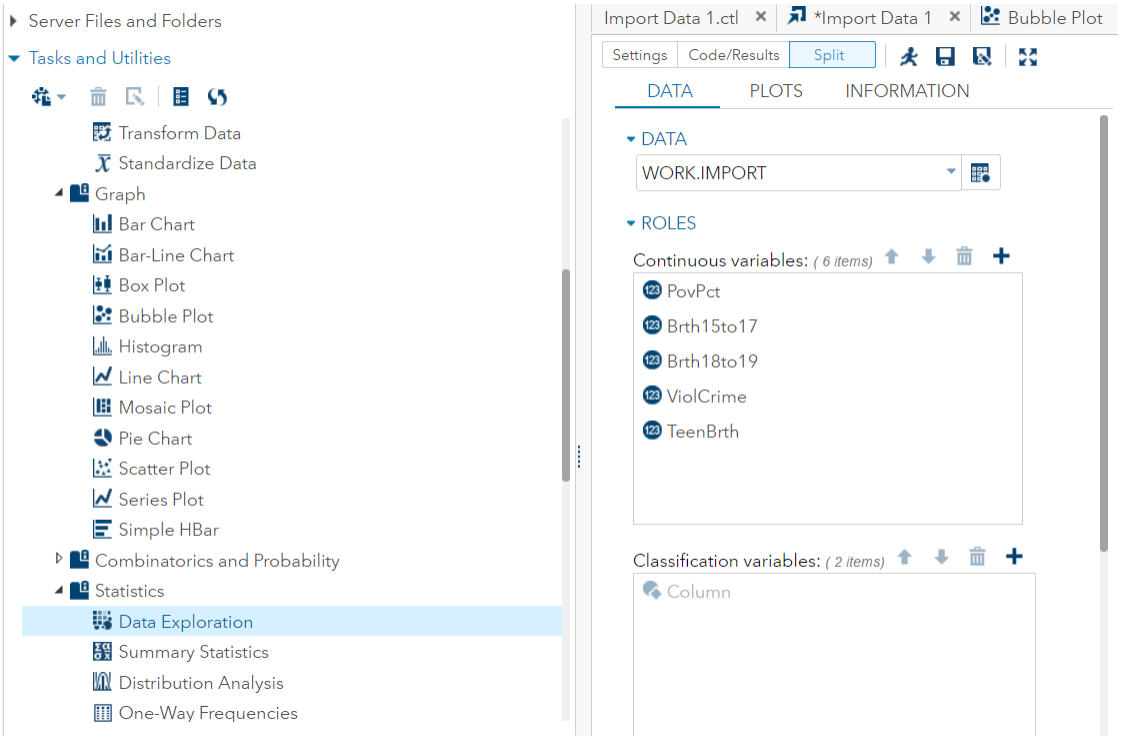

Distribution of ViolCrime
Percent
ViolCrime
Curves
Normal(Mu=7.8549 Sigma=8.9141)
Kernel(c=0.79)


Distribution of TeenBrth
Mean 42.243137255
Std Dev 12.318510541
Min 20
Max 69.1
Median 39.5
N 51
Q1 33
Q3 53
Q Range 20
Percent
TeenBrth

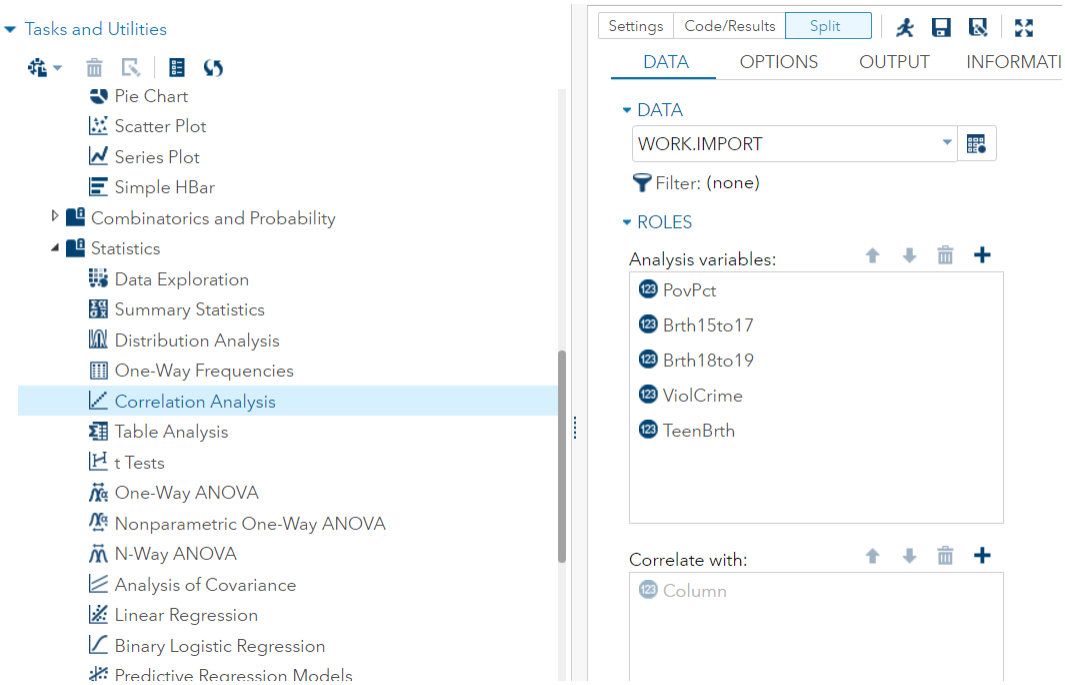

Distribution of TeenBrth
Percent
TeenBrth
Curves
Normal(Mu=42.243 Sigma=12.319)
Kernel(c=0.79)

2.3 Bivariate analysis

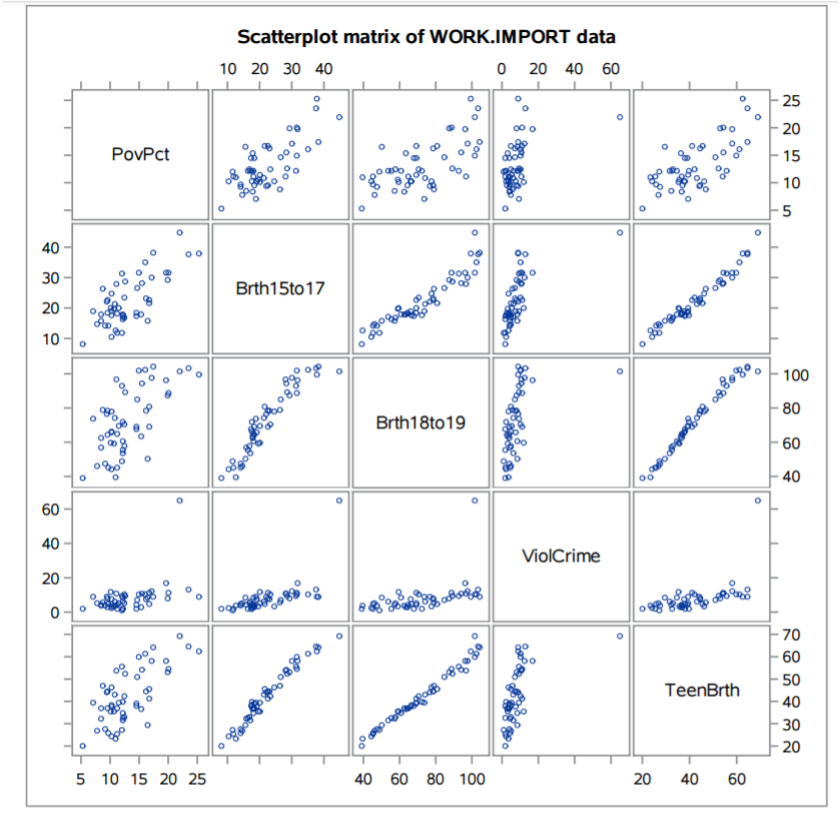
To analyse the dependencies between variables, it is useful to create a scatterplot matrix of the data and to compute the correlation matrix.
The scatterplot can be obtained by selecting the option **Data Exploration** under the window called **Statistics** .



The correlation matrix can be obtained by selection the option **Correlation Analysis** under the window called, **Statistics** .



2.4 Output



Looking at the scatterplot, all variables are positively correlated. Brth15to17, Brth18to19 and ViolCrime are highly correlated on each other. On the other hand ViolCrime is poorly

correlated with all variables. This is confirmed by the correlation matrix.

5 Variables:	PovPct Brth15to17 Brth18to19 ViolCrime TeenBrth

Pearson Correlation Coefficients, N = 51

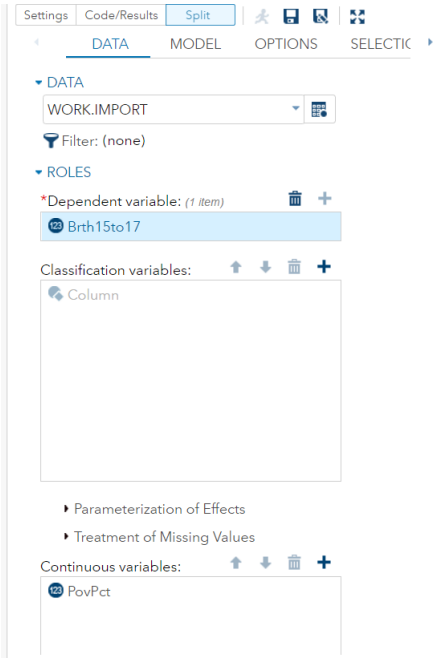
	PovPct	Brth15to17	Brth18to19	ViolCrime	TeenBrth
PovPct	1.00000	0.73029	0.64966	0.46956	0.70328
Brth15to17	0.73029	1.00000	0.94245	0.64027	0.97883
Brth18to19	0.64966	0.94245	1.00000	0.47770	0.98897
ViolCrime	0.46956	0.64027	0.47770	1.00000	0.55794
TeenBrth	0.70328	0.97883	0.98897	0.55794	1.00000

Indeed the highest correlations are 0.94245 (correlation between Brth15to17 and Brth18to19), 0.97833 (correlation between Brth15to17 and TeenBrth) and 0.98897 (correlation between Brth18to19 and TeenBrth). The lowest correlation is 0.46956 (correlation between PovPct and ViolCrime).

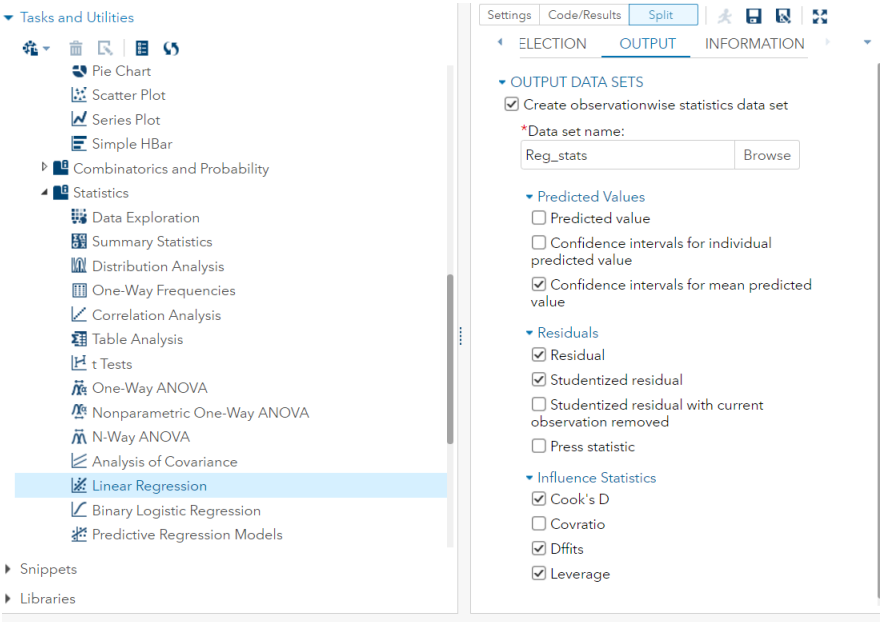
3 Linear Regression Model

We want to study the relationship between y = year 2002 birth rate per 1000 females 15 to 17 years old and x = poverty rate, which is the percent of the state's population living in households with incomes below the federally defined poverty level. Thus we focus on the scatterplot between PovPct and Brth15to17.

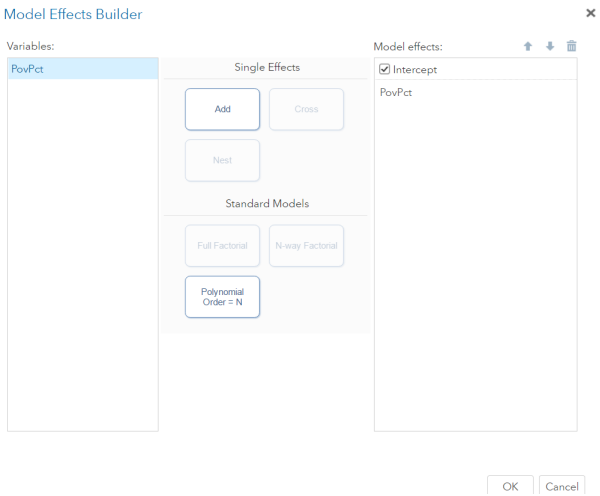
The plot shows a generally linear relationship, on average, with a positive slope. As the poverty level increases, the birth rate for 15 to 17 year old females tends to increase as well. We are now ready to estimate the linear regression model. To get the output select **Linear Regression** under the **Linear Models** window. Select Birth15to17 as Dependent variable and PovPct as Continuous variable.



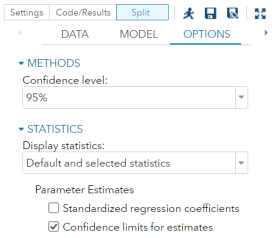
Open the Output window. Here there are some options, according to the aim of the analysis it is possible to compute the predicted values, to conduct a residual analysis or to investigate the presence of outliers and/or influential observations.



Then the model has to be built as follows, by opening the **Model** window. Add the variable PovPct in the Model effects column.



Finally, under the Option window, select the confidence intervals box for the estimates.



After selecting the run button, it is possible to get the estimates for the regression model.

3.1 Output

Model: MODEL1
Dependent Variable: Brth15to17

Number of Observations Read	51
Number of Observations Used	51

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1725.25949	1725.25949	56.00	<.0001
Error	49	1509.63463	30.80887		
Corrected Total	50	3234.89412			

Root MSE	5.55057	R-Square	0.5333
Dependent Mean	22.28235	Adj R-Sq	0.5238
Coeff Var	24.91018		

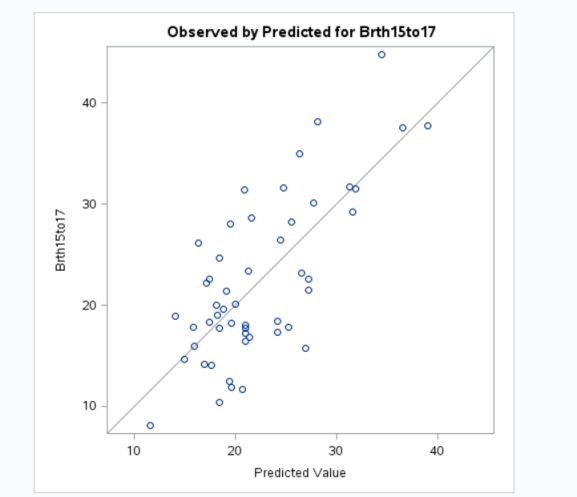
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	4.26729	2.52975	1.69	0.0980	-0.81642	9.35101
PovPct	1	1.37335	0.18352	7.48	<.0001	1.00454	1.74215

The estimated regression line is the following one $\hat{y} = 4.267 + 1.373 PovPct$, where $y = Brth150to17$

- The **interpretation of the slope** (value = 1.373) is that the 15 to 17 year old birth rate increases 1.373 units, on average, for each one unit (one percent) increase in the poverty rate.
- The **interpretation of the intercept** (value=4.267) is that if there were states with poverty rate = 0, the predicted average for the 15 to 17 year old birth rate would be 4.267 for those states. Since there are no states with poverty rate = 0 this interpretation of the intercept is not practically meaningful for this example.

From the output, we also see the information that $s = 5.55057$ (Root MSE) and $R^2 = 53.3\%$.



The value of s tells us roughly the average difference between the y-values of individual observations and predictions of y based on the regression line. The value of R^2 can be interpreted to mean that poverty rates explain 53.3% of the observed variation in the 15 to 17 year old average birth rates of the states. The R^2 (adj) value (52.4%) is an adjustment to R^2 based on the number of x-variables in the model (only one here) and the sample size. With only one x-variable, the adjusted R^2 is not important.

3.2 Simple Linear Regression Model Evaluation

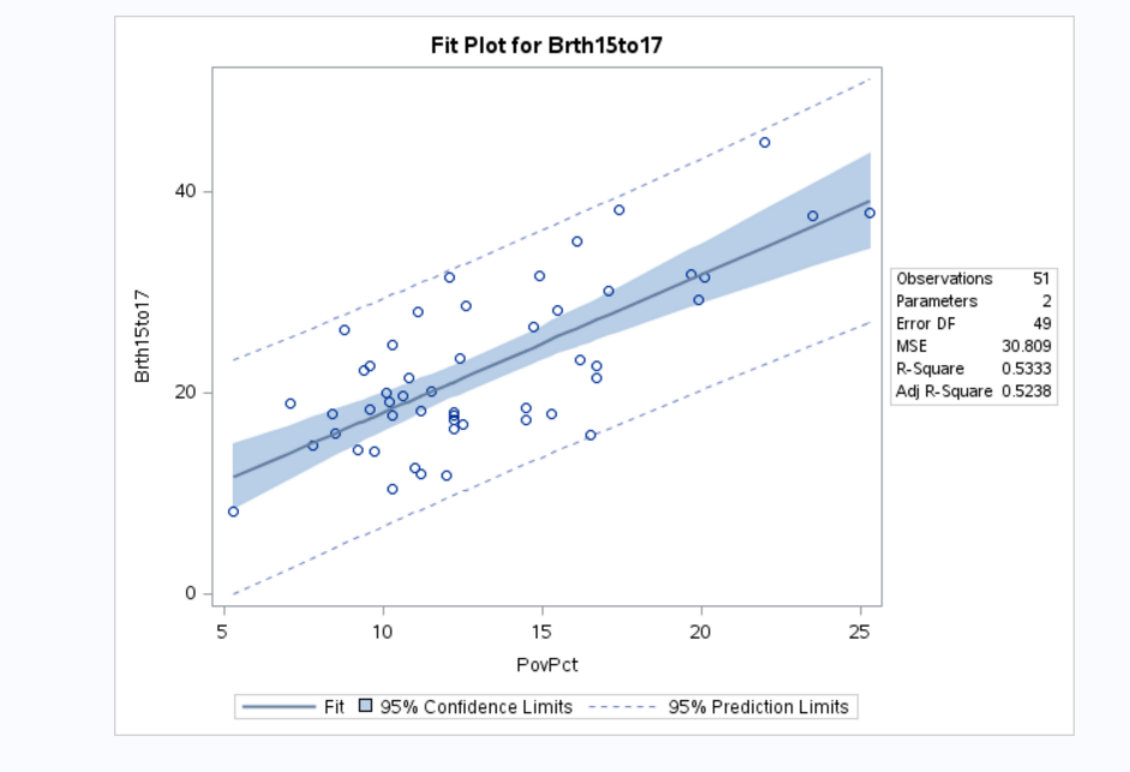
Recall that we are ultimately always interested in drawing conclusions about the population, not the particular sample we observed. In the simple regression setting, we are often interested in learning about the population intercept β_0 and the population slope β_1. Confidence intervals and hypothesis tests are two related, but different, ways of learning about the values of population parameters.
Is there a relationship between Brth15to17 and PovPct? Certainly, since the estimated slope of the line, $b1$, is 1.373, not 0, there is a relationship between Brth15to17 and PovPct in

the sample of 51 data points. But, we want to know if there is a relationship between the population of all of Brth15to17 and PovPct. That is, we want to know if the population slope β_1 is unlikely to be 0.

Looking at the p-value it is possible to conclude that β_1 is significantly different from 0 (p-value $<\alpha$, where α is usually 0.05 - thus the null hypothesis $H_0 : \beta_1 = 0$ is rejected).
This is equivalent to look at the confidence intervals; if 0 is included in the interval, then the corresponding β is not significant. It is possible to note that 0 is not included in the CI for β_1.
Furthermore it is very important to know the interpretation of the ANOVA table, and the meaning of all its elements (as shown in class).
The F-test is more useful for the multiple regression model when we want to test that more than one slope parameter is 0. Here F-test is equivalent to t-test (F-value=t-value^2 and the corresponding p-values are the same).

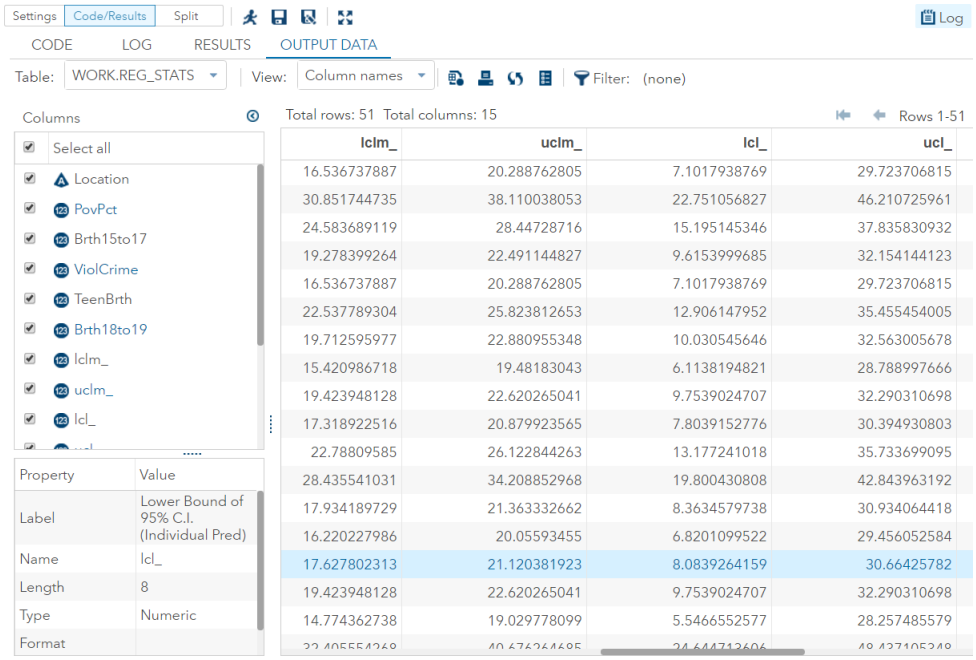
3.3 Simple Linear Regression Model Prediction

Typically, a regression analysis involves the following steps: model formulation, model estimation, model and model use. Here we focus on the following questions: what is the average response for a given value of the predictor $\mathbf{x}$? What is the value of the response likely to be for a given value of the predictor $\mathbf{x}$?

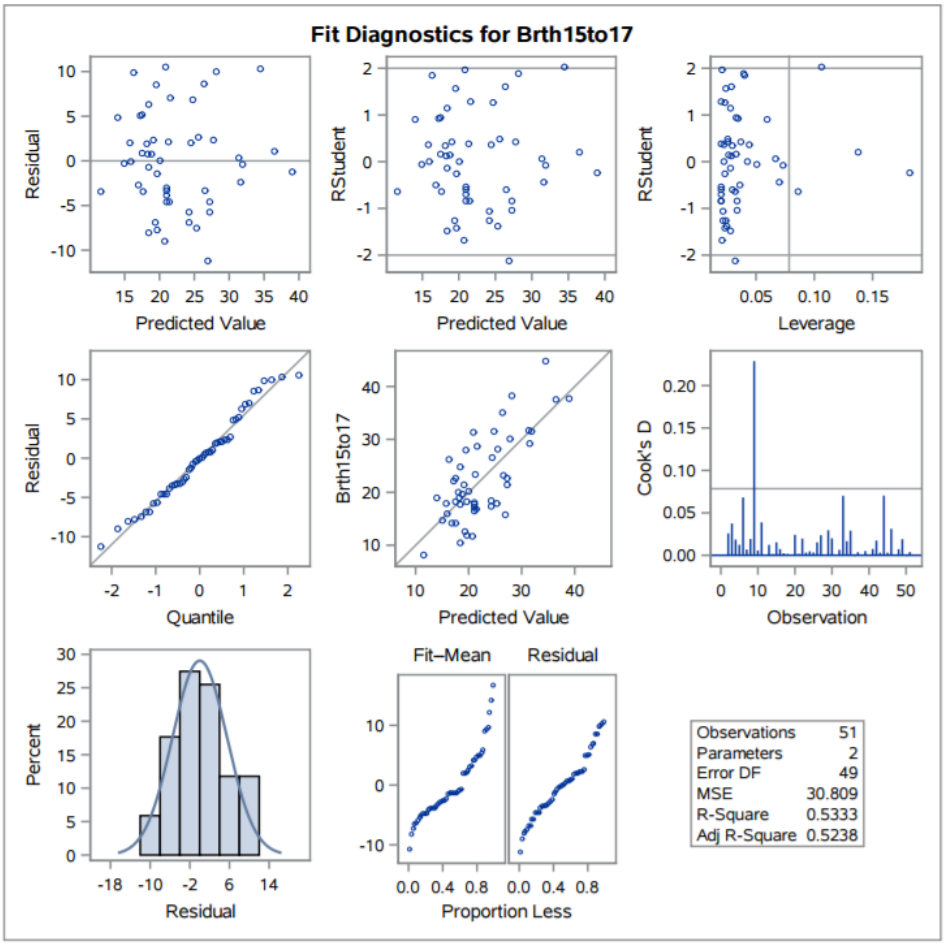


Observe that the prediction interval is always wider than the confidence interval. Furthermore, both intervals are narrowest at the mean of the predictor values (about 15).

(In SAS it is possible to get the specific CI or PI for a given x value, by selecting **Confidence intervals for individual predicted value** under the window Predicted Values). The output can be seen by selecting Output data window.



3.4 Simple Linear Regression Model Assumptions



Main Remarks

- **Residuals vs Predicted Value**. The residuals *bounce randomly* around the 0 line. This suggests that the assumption that the relationship is linear is reasonable. The residuals roughly form a *horizontal band* around the 0 line. This suggests that the variances of the error terms are equal. No one residual *stands out* from the basic random pattern of residuals. This suggests that there are no outliers.

- **Residuals vs Quantile**. Note that the relationship between the theoretical percentiles and the sample percentiles is approximately linear. Therefore, the normal probability plot of the residuals suggests that the error terms are indeed normally distributed. This is also confirmed by the histogram of the residuals.

- **Leverage and Cook's Distance**. They are useful to assess the presence of outliers and/or influential observations. Here only one observation needs further care to investigate its nature, that is observation number 9. (influential observation means that it could influence significantly the estimates of the regression line - graphically

the regression line is pulled up towards this observation. If the observation is not significantly, the estimates of the regression line with or without the observation under investigation are approximately the same. In other words, the regression line does not pull up towards this observation value).