Part 2. Modeling (Version 1.0 Date: 07033020)

Section 9: Simple Linear Regression

SalaryData.csv

Data science Files are located on GitHub:
Unix derivative users: clone -> https://GitHub.com/candy-o/DataScience.git
Windows users clone -> Download in Desktop requires "GitHub Desktop" or download Zip
(Mac O/S and Windows MSI)

- Here are my Section 49-54 notes following the video stream to guide us.

Google "Gretl" and install (Gretl.sourceforge.net (Windoze, Mac…)

Open source statistics program

*G*NU *R*egression, *E*conometrics and *T*ime-series *L*ibrary

SAS, IBM SPSS, R, Python similar

Use Notepad ++ then Libre Office calc to view SalaryData.csv

Then open in Gretl

…. Do you want to give the data a time series or panel interpretation? <No>

… Close parsing info window.

Have Variables and Constant

Go to view -> summary statistic
-Years experience

Then add in..
-Salary

Right click variable to extract values, edit values (not recommend), edit attributes…

Can save files as Gretl so that descriptions are not lost saving to .CSV

Select menu Model -> Ordinary Least Squares
Put Salary into dependent variable

Put YearsExperience in regressors
Const should be in regressors
Click <OK>

Model 1 report created…

First information regards our model file, type of regression OLS and variable(s) used.

Model 1: OLS, using observations 1-30
Dependent variable: Salary

Part 1. breaks down of statistics about the independent variables: Const and YearsExperience

Coefficient const line of linear regression crosses the 0 vertical axis at 25792.2 and each year salary goes up 9448.96 in the model world.

std. error, t-ratio, p-value but for this course we are talking about p-values mostly.

Pvalue allows us to filter out variables that are not going to predict anything. The smaller the Pvalue the better. Pvalue should be above threshold (0.50 5% default) so large const 5.51e-012 is not predicting anything, whereas small YearsExperience low 1.14e-020 is a good variable to use to predict.

Part 2
Here you have stats about the model most of all

Mean dependent Var
S.D. dependent Var
Sum squared resid
S.E. of regression

R-squared and Adjusted R-Squared are the most important.

F(1-28)
P-values(F)
Log-likelihood
Akaike criterion
Schwarz criteria
Hannan-Quinn

First way to bring up a chart in your model is Graphs -> Fitted, actual plot -> actual vs. Fitted

Analysis -> forecast ->static forecast -> click ok…

Two charts Graph and forecasts

95 percent chance interval salary is between green interval lines compared to not on the blue modeled line.