Developed by Elahe Rezaei

03/10/2019

In this report, I briefly explain how I fit linear model regression on five data sets. In order to derive a good estimator, first I take a look at each data set and their properties. I also look at distribution of y in each data set. As plots show, we will have data with different range of explanatories x, and number of samples. In some cases, like data set 1, there is a possibility of having outlier too. Although ordinary least squares (OLS) is said to be not robust for our non-normal data, we fit this model for comparison purpose. In addition, since the problem mention that the noise distribution depends on x, we may have dealing with heteroscedasticity. I manage the Regression Diagnostics and Specification Tests by following these tests:

- **Normality test**
- **Heteroscedasticity Test**
- **Outlier Detection**

I use three different techniques for finding the linear regression:

- **Generalized Least Squares (GLS)**
- **Formulating noise as distribution with variance dependent to x and applying Maximum Likelihood (MLE)**
- **Generalized Linear Model (GLM)**

These techniques can combine together in some cases, like combination of removing outlier with GLS. In Bayesian modeling, I pick two distributions with heavier tail: Student T and Half-Cauchy which are more usual for financial data. However, there are numerous different distributions which can be potentially a good candidate. I use package **pymc3** for Bayesian linear regression and package **scipy** for solving optimization problem with BFGS methods from Python.

The coding has been done in **Jupyter Notebook** and both the pdf format of Notebook and codes in Python are attached. The below table summarizes the performance of each method on data sets in terms of R^2. The final solution which contains the coefficients is attached as excel file.

	Set 1	Set 2	Set 3	Set 4	Set 5
OLS	0.164	0.398	0.103	0.028	0.281
GLS	0.185	0.420	0.103	0.025	0.276
OLS without Outlier	0.293	0.531	0.276	-	-
GLS without Outlier	0.293	0.5306	0.2765	-	-
MLE	0.291853	0.524300	0.191375	0.025596	0.2198
Student T	0.292537	0.48812	-1.3185	0.014900	-1.45905
Half-Cauchy	-	-	-	-	0.280563

Regression Diagnostics & Specification Tests

- **Normality Test**

After fitting a linear model OLS, I observe that the residues are non-gaussian. I perform Jarque - Bera normality tests. The normality test for linear model residues for the datasets lead to rejecting the null hypothesis that residue has a Gaussian distribution.

- **Heteroscedasticity Test**

For this test, the null hypothesis is that all observations have the same error variance, i.e. errors are homoscedastic. I employ het_white (Lagrange Multiplier Heteroscedasticity Test by White) by package statsmodels.

- **Outliers test**

These measures try to identify observations that are outliers, with large residual, or observations that have a large influence on the regression estimates. Robust Regression, RLM, can be used to both estimate in an outlier robust way as well as identify outlier. Based on my result, I decided to discard the one which is minimum weight in data_set [1], [2] and [3] as outlier and repeat the calculations without outlier for these data sets.

Linear Regression for Non-Normal Data

In our problem, the noise has heave-tailed distribution and dependent to explanatories. In order to avoid normality violation effect, nonparametric regression models are suggested. But it requires the large sample size. Thus ordinary least squares is said to be not robust to violations of its assumptions and we exploit Generalized Least Squares.

- **Generalized Least Squares**

As the data is heteroskedastic, we assume that the error terms follow an AR(1) process with a trend:

$$\epsilon_i = \beta_0 + \rho\epsilon_{i-1} + \eta_i$$

where $\eta \sim N(0, \sigma^2)$ and that ρ is simply the correlation of the residual a consistent estimator for ρ is to regress the residuals on the lagged residuals. The package statsmodels provide variations of GLS.

- **Formulating The Problem and MLE Estimator**

In this section, I will propose a distribution for noise which is dependent to explanatories (x). Then, by using package scipy from Python, I optimize the likelihood function and find the parameters of model (a, b) and noise distribution statistics. For simplicity, I consider noise as $\epsilon_i \sim N\,(0, \sigma_0^2) + N\left(0, x_i^2\sigma_1^2\right)$ where σ_0 and σ_1 will be obtained through the optimization of log-likelihood function. We can consider other distributions like T-distribution and see which one will yield the best result.

$$y_i = \beta x_i + \alpha + \epsilon_i$$
$$where\ \epsilon_i \sim N\,(0, \sigma_0^2) + N\left(0, x_i^2\sigma_1^2\right)$$

$$p\left(Y\middle|X = x, a, b, \sigma_i^2\right) = \prod_{i=1}^{N} p\left(y_i\middle|x_i, a, b, \sigma_i^2\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(-\frac{(y_i - ax_i - b)^2}{2\sigma_i^2}$$

Where $\sigma_i^2 = \sigma_0^2 + x_i^2\sigma_1^2$. So the negative log-likelihood function is

$$NLL = \sum\nolimits_{i=1}^{N} \frac{1}{2}\log\left(\sigma_i^2\right) + \frac{(y_i - ax_i - b)^2}{2\sigma_i^2}$$

In order to find MLE estimation, we will solve the following set of equations:

$$\frac{\partial NLL}{\partial a} = \sum\nolimits_{i=1}^{N} \frac{-2x_i(y_i - ax_i - b)}{2\sigma_i^2}$$

$$\frac{\partial NLL}{\partial b} = \sum\nolimits_{i=1}^{N} \frac{-2(y_i - ax_i - b)}{2\sigma_i^2}$$

$$\frac{\partial NLL}{\partial \sigma_0} = \sum\nolimits_{i=1}^{N} \frac{\sigma_0}{\sigma_i^2} + \sum\nolimits_{i=1}^{N} \frac{-(y_i - ax_i - b)^2\sigma_0}{\sigma_i^4}$$

$$\frac{\partial NLL}{\partial \sigma_1} = \sum\nolimits_{i=1}^{N} \frac{\sigma_1 x_i^2}{\sigma_i^2} + \sum\nolimits_{i=1}^{N} \frac{-(y_i - ax_i - b)^2\sigma_1\, x_i^2}{\sigma_i^4}$$

- **Generalized Linear Model (GLM)**

Here I fit a Bayesian linear regression model to the data. The model specifications in PyMC3 are wrapped in a with statement. Here we use the NUTS sampler to draw 3000 posterior samples.

I find a robust regression model by choosing an error distribution with fatter tails; a common choices are Student's t-distribution and half-Cauchy distribution.

It is worth to mention that there are many other methods which are recommended for regression while non-normal data, such as *data transform* and table 1 can be expanded.