A Robust Study of Regression Methods for Crop Yield Data

Ying Zhu
SAS Institute, Inc
ying.zhu@sas.com

Sujit K. Ghosh
Department of Statistics
North Carolina State University

Selected Paper prepared for presentation at the Agricultural & Applied Economics Associations 2011 AAEA & NAREA Joint Annual Meeting, Pittsburgh, Pennsylvania, July 24- 26, 2011.

A Robust Study of Regression Methods for Crop Yield Data

Ying Zhu* SujitK.Ghosh**

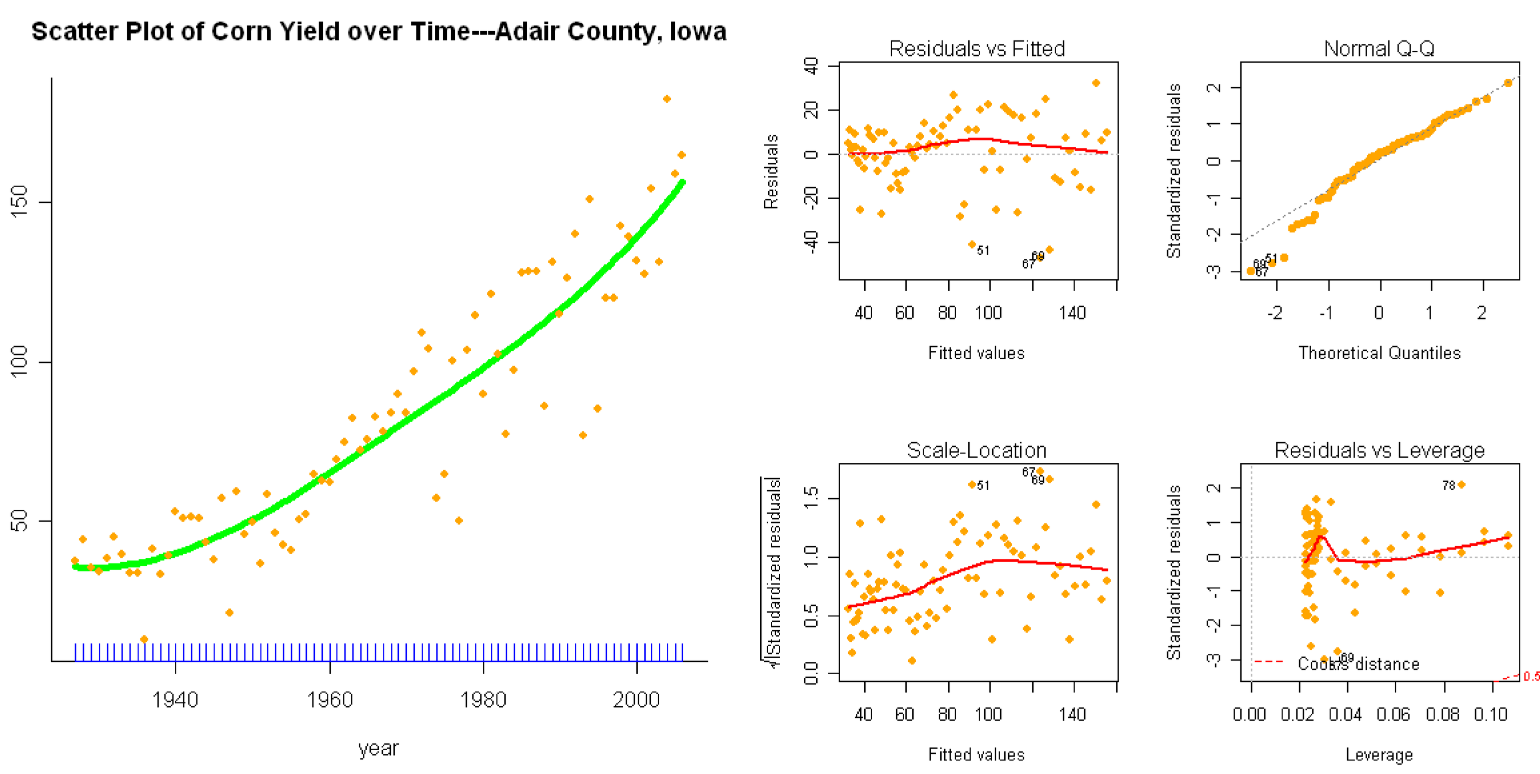
* SAS Institute, Inc, Cary, NC ** North Carolina State University, Raleigh, NC

Presented at AAEA 2011, Pittsburgh, Pennsylvania, July 24-26, 2011

OBJECTIVE

- Precisely estimating crop yield risk is crucial for the proper design and rating of crop insurance contracts
- Advances in biotechnology and changes in environmental conditions may significantly affect the distributions of crop yields
- These changes can complicate efforts to accurately model yield distributions using time series data
- The objective of this study is to evaluate the robust regression methods when detrending the crop yield data. We analyze the properties of robust estimators for outliers contaminated data in both symmetric and skewed distribution case.

MOTIVATION



The figure above gives the plot of the county-level corn yield. It shows:

- Upward time trend—we need to estimate/remove the trend when model yield distribution
- Outliers: Outliers can shift trend estimation arbitrarily far from the real
- Skewed: Left-skewed from Q-Q plots—an indication for non-normality
- Heteroscedastic: Non-constant coefficient variance—errors are proportional to mean

A SIMPLE DETRENDING REGRESSION

Consider the following trend model:

$$y_t = \beta_0 + \beta_1 t + \epsilon_t$$

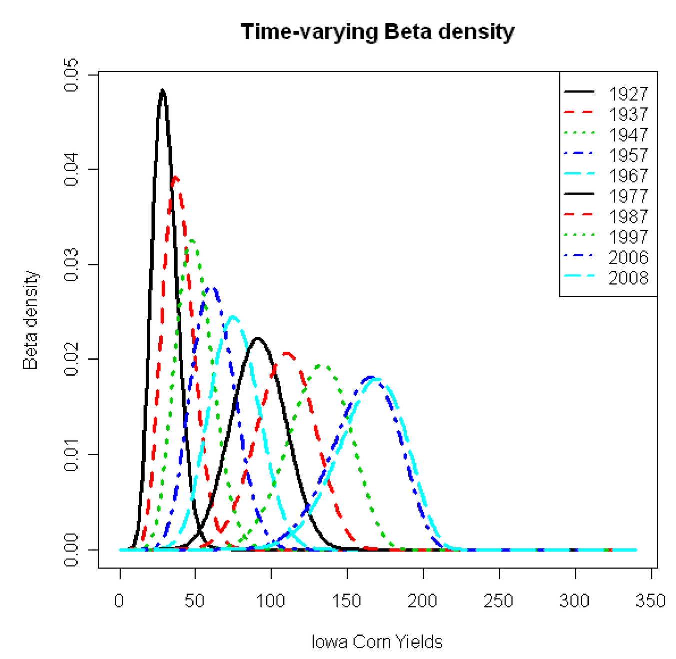
- y_t is the observed crop yield in year t, $(t = 1, \ldots, T)$,
- ϵ_t represent residuals that are assumed to be independently distributed with mean zero: $E(\epsilon_t) = 0$ and $var(\epsilon_t) = \sigma_t^2$

THE ROBUST REGRESSION METHODS

We consider the following robust regression techniques for the detrending of crop yield data:

- OLS
- M estimation introduced by Huber (1973)
- MM estimation introduced by Yohai (1987)
- Time-varying Beta method by Zhu, Goodwin and Ghosh (2011)

Time-varying Beta density plots based on MLE estimates



- Above figure shows time-varying Beta density plot: $y_t \sim Beta(\alpha_t, \beta_t, \theta_t, \delta_t)$
- The density plots show that various moments of the distributions evolve over time as the technology progresses

MONTE CARLO SIMULATIONS

- The Monte Carlo simulation is used to study the performance of the candidate robust regression method.
- The simulation parameter is chosen to be consistent with the robust study in the previous literatures (Swinton and King (1991))
- Fix β_0, β_1 for some positive numbers
- Yield series are generated assuming a known trend of β_1 and a random error with variance σ_t^2

Simulation Under Symmetric and Skewed Distributions

- Symmetric Normal Distribution: $\epsilon_t \sim N(\mu = 0, \sigma = \sigma_t)$
- Skewed Beta Distribution: $\epsilon_t \sim 6\sigma_t(Beta(10,6) - \frac{5}{8})$
- Set $\sigma_t = 25t^{\alpha}$, where $\alpha \in [0, 1)$.
- This variance form introduces a general outlier generating form when α equals to any nonzero number. The largest variance will occur at the end of the series
- Both the distributions are designed properly so that $\sigma_t = 125$ when $t = T$
- Set $T \in \{10, 15, 20, 25\}$ and $\alpha \in \{0, 0.1, 0.2, \frac{\log 5}{\log T}\}$
- Different value of sample size T and α stands for different outlier contamination scenarios

EMPIRICAL RESULTS

- 1000 datasets are generated under each distribution with different value of α and T.
- The identical simulated yield series are fitted using OLS, M estimator, MM estimator, time-varying Beta models

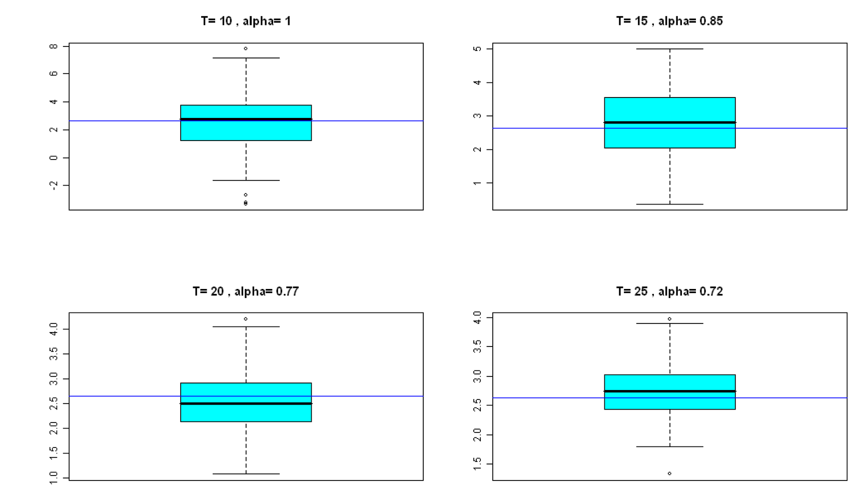


Figure: Box Plots—Normal Error Terms

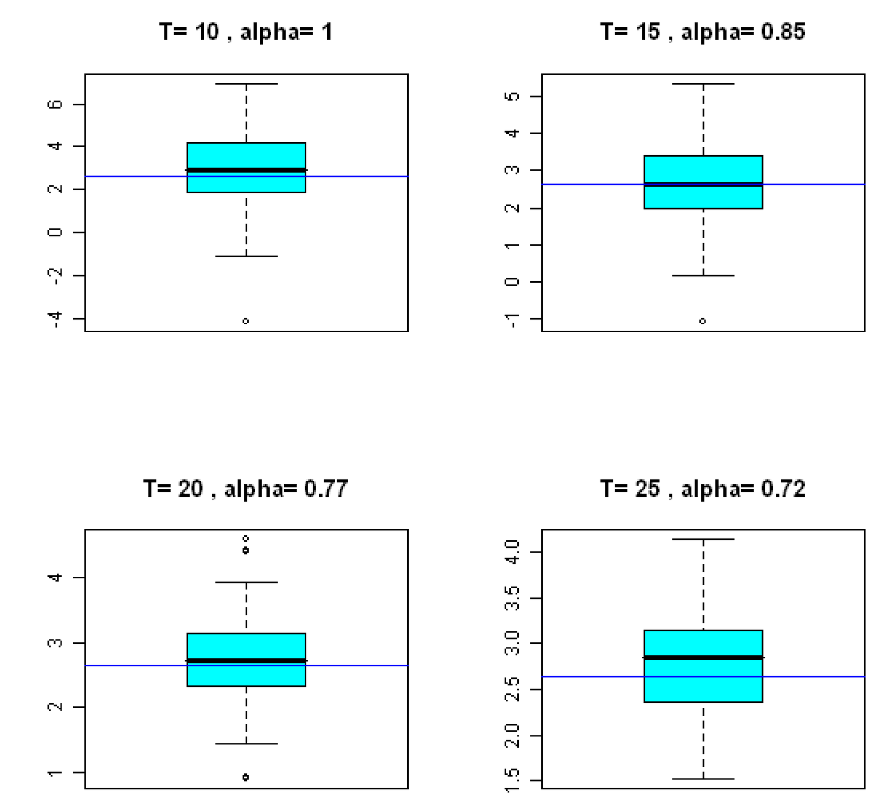


Figure: Box Plots—Beta Error Terms

- Above figure shows an example of box plots for OLS estimates for each distribution and sample size

MODEL PERFORMANCE AND ECONOMICS IMPLICATION

- Boxplots of OLS, M estimates, MM estimates and time-varying Beta estimates allow us to compare the in-sample robustness of these regression methods
- The performance of the candidate robust estimators is assessed in terms of trend estimation, out-of-sample prediction of future yield levels using the simulated series
- Implication of crop insurance rate estimation is analyzed based on different estimation methods