Statistical Models for Extreme Values

Peter Julian A. Cayton
Assistant Professor, University of the Philippines Diliman

Questions on the extreme scenarios in environmental conditions and economic climate have surfaced over the past years. Issues on global climate change have brought about questions on possible severe weather conditions such as droughts, floods, and heavy rains, which result to loss of lives and livelihood. Financial and economic crises have introduced uncertainties in economic conditions that give rise to possibilities of bankruptcies in global institutions and high inflation levels which could bring serious threat to society. Extreme scenarios are rare, yet the effects persist in very long time frames and impacts diverse scenarios. Statistical methods depicting typical central values of observed variables could fail in describing extreme conditions. Thus, extreme value theory offers alternative methods in understanding extreme behavior.

What is Extreme Value Theory?

Extreme value theory offers results in distribution theory that can help characterize the behavior of extreme observations. Analogous to central limit theory (describes how the sample mean converges to the normal or Gaussian distribution as sample size gets very large), extreme value theory describes how the sample maximum or the tail of a distribution converges to a probability distribution as sample size gets very large. The central limit theory favors statistical inference through the Gaussian distribution, e.g., characterization of the sample mean, sample variance, conduct of t-tests, z-tests, and F-tests, etc. On the other hand, extreme value theory is associated with estimation of extreme quantities. The literature however is dominated by topics on modeling sample maxima than sample minima values. This does not pose any dislike to modeling extremely low values, but it is more brief to discuss one type of extreme value since a minimum can be easily transformed to a maximum value without loss of information, i.e., $\max(-X_1,\ldots,-X_n) = -\min(X_1,\ldots,X_n)$.

Two Theorems and Some Methodologies of Extreme Value Theory

Given a random sample $X_1,...,X_n$, from a population distribution function F_x, the distribution of the sample maximum, $X_{(n)}$ is $F_{x(n)}(x) = [F_x(x)]^n$, which would degenerate when the sample size n becomes very large (Mood et al., 1974). This result is trivial since it implies that inference on the maximum value is futile for very large sample sizes. However, convergence can be weakened as sample size becomes large to generate elegant results for extreme values.

The first theorem that describes the property of extreme values is the Fisher-Tippett-Gnedenko theorem (Fisher and Tippett, 1928), also known as the Extremal Types Theorem. If a sequence of values a_n and b_n is properly chosen, both as functions of n, then the sequence $z_n = (X_{(n)} - a_n)/b_n$ converges as n becomes very large to one of three distributions depending on certain criteria on the population distribution. The theorem is modified (Coles, 2001) so that the convergence reduces to one distribution, the generalized extreme value (GEVD) distribution, which summarizes the three distributions. The GEVD distribution is shown having a distribution function $G(x; \mu,\sigma,\xi)$ of the form:

$$G(x) = \begin{cases} \exp\left\{-\left[1+\xi\left(\dfrac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, & \text{for } 1+\xi\left(\dfrac{x-\mu}{\sigma}\right) \geq 0 \text{ and } \xi \neq 0 \\ \exp\left\{-\exp\left[\dfrac{x-\mu}{\sigma}\right]\right\} & \text{for } -\infty < x < +\infty \text{ and } \xi = 0. \end{cases}$$

Based on this distribution, high quantiles, called return levels, can be derived. An m-period return level is the $1/m$ upper quantile of the distribution, interpreted as the "once in every m observations/periods" extreme value. Under the GEVD, this return level is:

$$x_m = \begin{cases} \mu - \dfrac{\sigma}{\xi}\left[1-\left(\ln\dfrac{m}{m-1}\right)^{-\xi}\right] & \text{for } \xi \neq 0 \\ \mu - \sigma\left[\ln\left(\ln\dfrac{m}{m-1}\right)\right] & \text{for } \xi = 0 \end{cases}$$

The data fitted to the GEVD are maximum values of a characteristic observed at discrete time points (e.g., annual maximum sea level of a bay, daily maximum wind speed, maximum day-by-day percentage loss of investing in a market asset in a month). Estimation of the parameters of the GEVD may be carried out through any manner of estimation, e.g., maximum likelihood (Coles, 2001),

method of moments (de Haan and Ferreira, 2006), or Bayesian techniques (Coles and Davison, 2008). From the estimated parameter values, the m-period return level can be estimated by substitution.

When the dataset $\{r_1,...,r_T\}$ is not a collection maximum value at discrete time points (e.g., rate of asset return in one day of market trading), the data is divided into g sub-samples of consecutive periods called blocks of equal size M such as months, or quarters, i.e., $\{[r_1, ...,r_M],..., [r_{(g-1)M+1}, ...,r_{gM}]\}$. From each block, the maxima $r^*_j = r_{(j-1)M+1},...,r_{jM}$ are collected and the maxima dataset $\{r^*_1,...r^*_g\}$ are fitted to the GEV distribution. This methodology is called the "block maxima" procedure (Gilleland and Katz, 2005). One typical problem with the block maxima procedure is its high sample size requirement. A large number of blocks are necessary for optimal inference with the GEVD. Within each block, a large number of time points should be contained to provide an appropriate maximum value for the block, e.g., a maximum for a 5-day trading week is less reliable compared to a maximum for a 20-day trading month for daily data.

The second theorem that describes extreme occurrences is the Pickands-Balkema-de Haan theorem (Balkema and de Haan, 1974; Pickands, 1975), which is the basis of the "peaks-over-thresholds" approach in extreme value models. The theorem states that for a very large threshold value u, the conditional probability distribution $P(X > x \mid X > u)$ of a random variable X approaches the generalized Pareto distribution (GPD), with distribution function $H(x; \mu,\sigma,\xi)$ and m-period return level x_m (Coles, 2001):

$$H\left(x;u,\sigma,\xi\right)=\begin{cases}1-\left(1+\xi\dfrac{x-u}{\sigma}\right)^{-1/\xi} & for\ \ \xi\neq 0\ and\ x>u\\ 1-\exp\left(-\dfrac{x-u}{\sigma}\right) & for\ \ \xi=0\ and\ x>u\end{cases}$$

$$x_m=\begin{cases}u+\dfrac{\sigma}{\xi}\left[\left(m\zeta_u\right)^{\xi}-1\right] & for\ \xi\neq 0\\ u+\sigma\ln\left(m\zeta_u\right) & for\ \xi=0.\end{cases}$$

The term ζ_u is equal to $P(X > u)$. Given some data, the return level is estimated by the proportion of observations greater than u in the sample.

To carry out the peaks-over-thresholds approach on a dataset $\{r_1,...,r_T\}$, a proper threshold $\tilde{u}$ is selected first. Some threshold selection methods are compiled in Coles (2001) and Coles and Davison (2008), yet expert opinions can be used as basis. From the dataset, all observations whose values are less than or equal to $\tilde{u}$ are ignored in the estimation process and those which are larger than

$\tilde{u}$ are gathered to a new dataset $\left\{r_1^{\tilde{u}},...,r_d^{\tilde{u}}\right\}$, where d is the size of the new dataset. Estimation of the parameters of the GPD is carried out using the new dataset using any method of estimation. This is then used in estimation of return levels $\hat{x}_m$ with $\hat{\zeta}_{\tilde{u}} = \left(T-d\right)/T$.

The difficulty of using the method is the selection of the threshold, which is analogous to finding the "goldilocks zone." If the threshold is too small, then there is a problem of bias since the collected data may not portray the true tails of the data, i.e., some typical values are mixed in with the extremes. If the threshold is too large, then a very small number of observations will be used for GPD estimation, and thus leads to unreliable parameter estimates.

Coles (2001) and de Haan and Ferreira (2006) provide more comprehensive discussion on different scenarios in theory and applications in extreme value theory.

Applications of Extreme Value Theory

There have been some promising applications of extreme value statistics in the environmental sciences. Return levels on certain environmental variables bring insight on possible scales of severe conditions, such as the "once-every-100-year flood level," "once-every-decade storm surge," and the like as basis for structural standards for flood gates and seas walls. Coles (2001) compiles these applications in environmetrics and Castillo et al. (2005) compiles its applications in science and engineering, where structural integrity studies, quality control, and other examples are shown.

Recent applications of extreme value theory has been observed in the literature of econometrics. By the adoption of the Basel accords by central banks and financial regulators around the world, value-at-risk estimation by financial institutions through statistical modeling has flourished, and the "100-period return levels" in extreme value theory can be interpreted as "once-every-100-trading maximum loss" or the value-at-risk at 1% probability of extreme loss. Jorion (2007), Tsay (2002), and McNeil et al. (2005) documented the use of extreme value theory in financial econometrics and market risk management and derive the value-at-risk measure based on return levels.

Among the early research in extreme value statistics in the Philippines, Formacion et al. (1991) applied extreme value theory to explain the statistical distribution of the largest fish lengths of mackerel. More recently, econometric applications were contained in Suaiso and Mapa (2009) and Cayton et al. (2010),

where extreme value theory is used in value-at-risk estimation of some financial indicators. Santos et al. (2010) discusses the use of extreme value theory in price level monitoring through the inflation-at-risk measure.

Final Remarks

Answering questions concerning potential extreme scenarios and magnitudes through data is possible through statistical modeling of extreme values based on extreme value theory. The methodologies have strong applications in environmental science and financial engineering where extreme scenarios are possible to cause adverse effects. Extreme value statistics and its applications in the Philippines are open research frontiers with the potential of making significant contributions in environmental and economic planning, monitoring, and policymaking.

References

BALKEMA, A., and L. DE HAAN, 1974, Residual life time at great age, *Annals of Probability*, 2, 792–804.

CASTILLO, E, A.S. HADI, N. BALAKRISHNAN and J.M. SARABIA, 2005, *Extreme Value and Related Models with Applications in Engineering and Science*, Hoboken, New Jersey: John Wiley & Sons. Inc.

CAYTON, P.J.A., C.D.S. MAPA and M.T. LISING, 2010, Estimating Value-at-Risk (VaR) Using TiVEx-POT Models, *Journal of Advanced Studies in Finance*, December 2010.

COLES, S., 2001, *An Introduction to Statistical Modeling of Extreme Values*, London: Springer-Verlag.

COLES, S., and A. DAVISON, 2008, *Statistical Modeling of Extreme Values*, Internal Talks of Competence Cetner Environment and Sustainability, A Portable Document File (PDF) downloaded at March 21, 2012 12:20am, at https://edit.ethz.ch/cces/projects/hazri/EXTREMES/talks/colesDavisonDavosJan08.pdf.

DE HAAN, L. and A. FERREIRA, 2006, *Extreme Value Theory: An Introduction*. New York, NY: Springer Science+Business Media, LLC.

FISHER, R.A. and L.H.C. TIPPETT, 1928, "Limiting forms of the frequency distribution of the largest or smallest member of a sample," *Mathematical Proceedings of the Cambridge Philosophical Society*, 24, pp 180-190.

FORMACION, S.P., J.M. RONGO and V.C. SALIMBAY, 1991, "Extreme Value Theory Applied to the Statistical Distribution of the Largest Lengths of Fish," *Asian Fisheries Science*, Vol. 4: 123-135.

GILLELAND, E., and R. KATZ, 2005, Extremes Toolkit (extremes): Weather and Climate Applications of Extreme Value Statistics. A Portable Document File (pdf) accessed March 10, 2009 from http://www.isse.ucar.edu/extremevalues/evtk.html.

JORION, P., 2007, *Value at Risk: The New Benchmark for Managing Financial Risk, 3rd Ed.*, USA: McGraw-Hill.

McNEIL, A.J., R. FREY, and P. EMBRECHTS, 2005, *Quantitative Risk Management: Concepts, Techniques and Tools*. USA: Princeton University Press.

MOOD, A. M., F.A. GRAYBILL and D.C. BOES, 1974, *Introduction to the Theory of Statistics*, 3rd Ed. USA: McGraw-Hill.

PICKANDS, J., 1975, "Statistical inference using extreme order statistics," *Annals of Statistics*, 3, 119–131.

SANTOS, E.P., C.D.S. MAPA, and E.T. GLINDRO, 2010, "Estimating Inflation-at-Risk (IAR) Using Extreme Value Theory (EVT)," Invited Paper Session of the 11th National Convention in Statistics, October 4-5, 2010.

SUAISO, J.O.Q. and C.D.S. MAPA, 2009, "Measuring Market Risk Using Extreme Value Theory," *Philippine Review of Economics*, Vol. 46, No. 2.

TSAY, R.S., 2002, *Analysis of Financial Time Series*, USA: John Wiley & Sons.