Robust Statistics and Financial Data

Colin T. Bowers

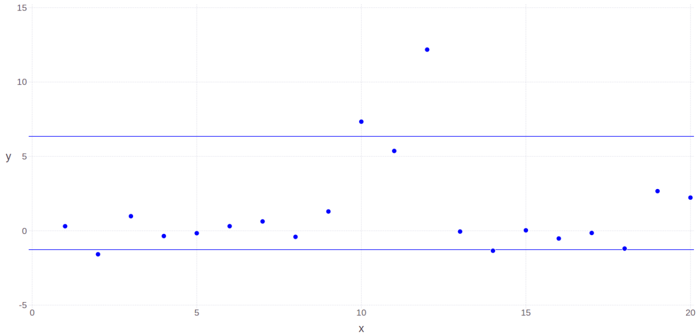
Macquarie University,
colintbowers@gmail.com,
https://github.com/colintbowers/RobustStatsTutorial.jl

2016-09-14

Preview

1. Basic intuition and the trimmed mean
2. Measuring tail fatness
3. Fat tails in financial returns
4. The trimmed mean and financial returns
5. Robust estimation of a linear model for financial returns
6. Traps and pitfalls when using robust statistics

Fat-tailed data

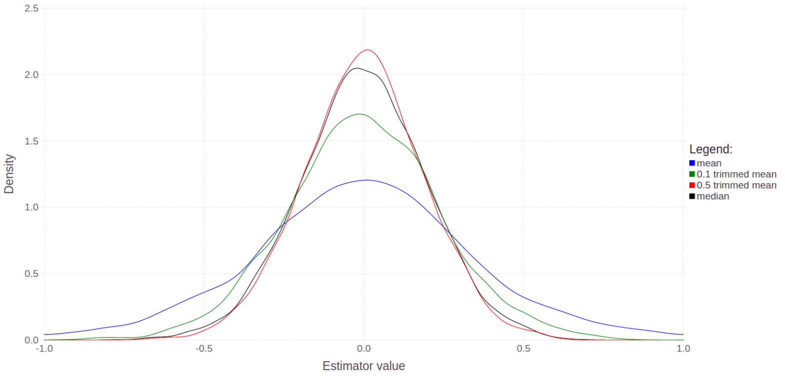


20 observations from the Student-t (2 DoF*) Distribution

- True mean = 0
- Sample mean = 1.38
- Trimmed mean = 0.68

*DoF = Degrees of Freedom

The trimmed mean



Simulated kernel density estimate for the trimmed mean given a Student-t (2 DoF) DGP*

*DGP = Data Generating Process

Measures of tail fatness

Standard textbook definition:

$$Kurtosis = \frac{\mathbb{E}(X-\mu)^4}{(\mathbb{V}X)^2} \tag{1}$$

Problem: What if $\mathbb{E}X^4 \to \infty$?

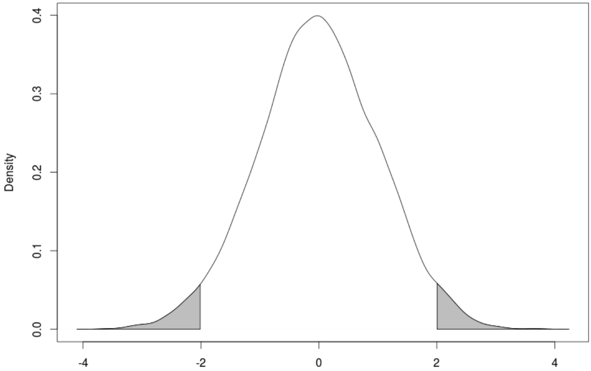
Solution: Hogg's robust kurtosis:

$$RobustKurtosis = \frac{\mathbb{E}(X\mathbb{I}\{X > Q_{0.95}\}) - \mathbb{E}(X\mathbb{I}\{X < Q_{0.05}\})}{\mathbb{E}(X\mathbb{I}\{X > Q_{0.5}\}) - \mathbb{E}(X\mathbb{I}\{X < Q_{0.5}\})} \tag{2}$$

where Q_p is the quantile associated with probability p.

Hogg's robust kurtosis numerator

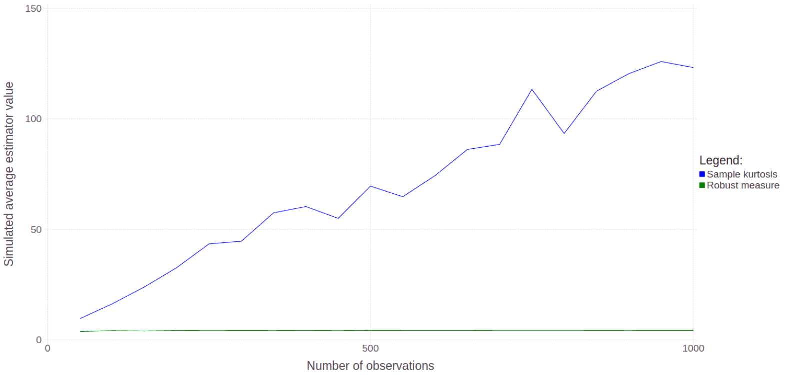
Numerator: $\mathbb{E}(X\mathbb{I}\{X > Q_{0.95}\}) - \mathbb{E}(X\mathbb{I}\{X < Q_{0.05}\})$



Hogg's robust kurtosis denominator

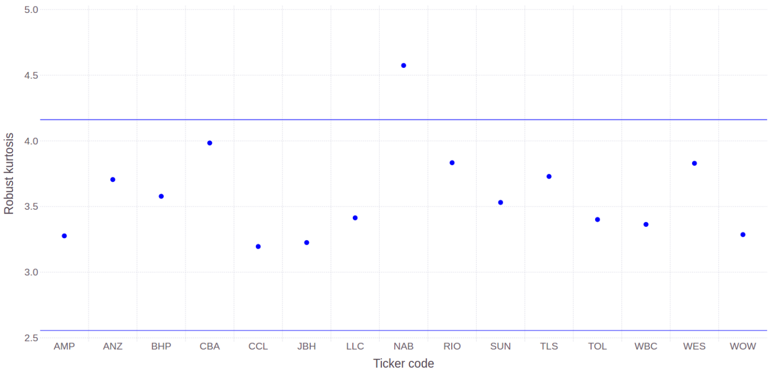
Denominator: $\mathbb{E}(X\mathbb{I}\{X > Q_{0.5}\}) - \mathbb{E}(X\mathbb{I}\{X < Q_{0.5}\})$

The problem with undefined moments



Sample kurtosis versus robust kurtosis for Student-t (2 DoF) DGP

Tail Fatness in Daily Financial Returns



Robust kurtosis of daily financial returns for some popular stocks. Note, Normal and Student-t (2 DoF) lines included for reference.

A model for unconditional fat tails

Unconditional fat tails can be generated by the model:

$$r_t \backsim \mathcal{N}(\mu, \sigma_t^2), \tag{3}$$

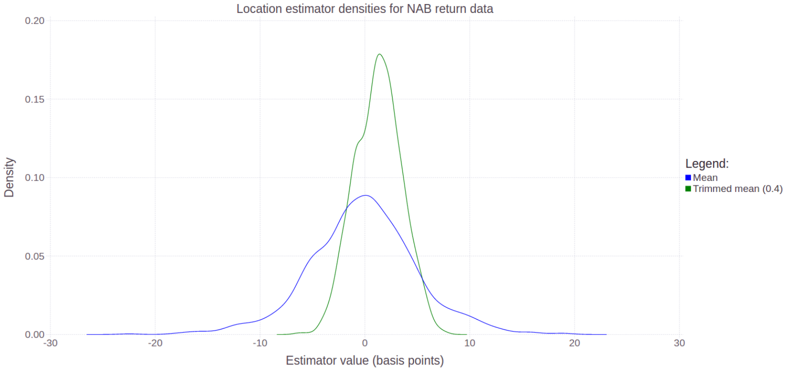
where σ_t is typically stochastic, sometimes by conditioning on time $t-1$ (e.g. GARCH).

The volatility clustering effect of GARCH results in short periods where, *unconditionally*, we get many observations from the tails, e.g. late 2008 to early 2009.

Resampling financial returns

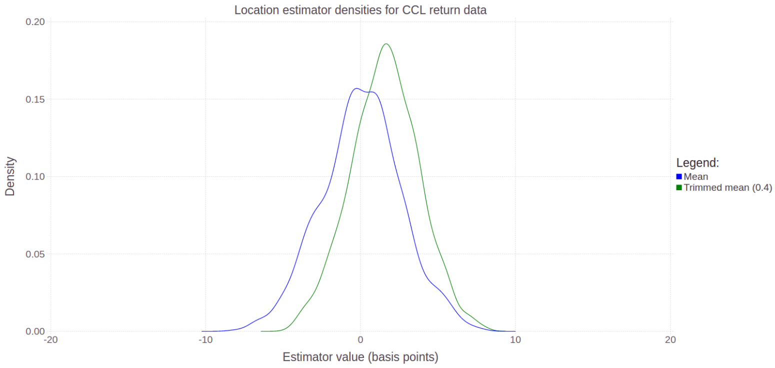
- Start with a sequence of returns $r_t, t = 1, ..., T$
- Centre the returns, i.e. $z_t = r_t - \bar{r}$
- Sample (with replacement) blocks of observations from $z_t, t = 1, ..., T$. Denote a re-sampled observation z_t^*.
- Under fairly general assumptions, $\mathbb{E}z_t^* = 0$, but z_t will otherwise have "similar" statistical properties to r_t

Kernel density estimates for trimmed and sample mean



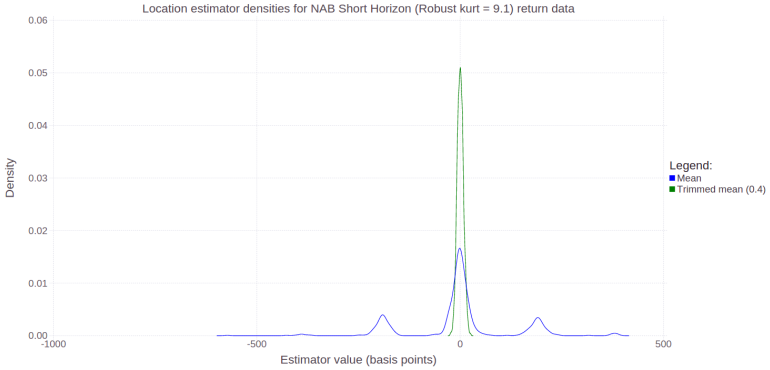
Resampled NAB daily return data (2005 to 2015).

Kernel density estimates for trimmed and sample mean



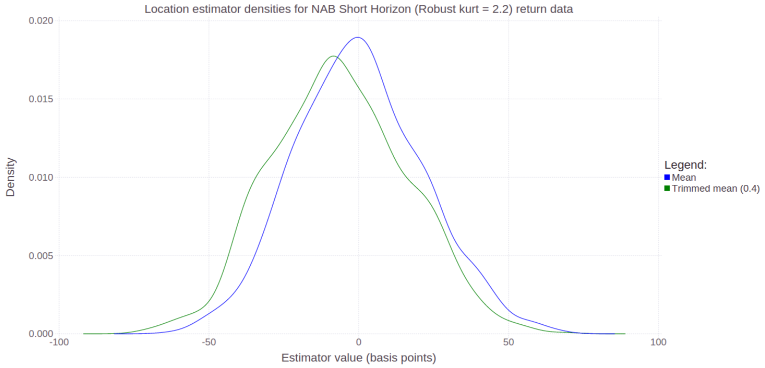
Resampled CCL daily return data (2005 to 2015).

Kernel density estimates for trimmed and sample mean



Resampled NAB daily return data (100 sequential days with largest robust kurtosis).

Kernel density estimates for trimmed and sample mean



Resampled NAB daily return data (100 sequential days with smallest robust kurtosis).

A univariate linear model

Assume daily returns are generated by the model:

$$r_t = \alpha + \beta s_t + e_t, \tag{4}$$

where s_t is a predictive signal.

Note: If the R^2 of this model is small, then r_t and e_t are likely to have similar statistical properties, such as heteroskedasticity.

A broad class of estimators for a linear model

Given observable data vectors $\mathbf{r}$ and $\mathbf{s}$, we have $\mathbf{e} = \mathbf{r} - \alpha - \beta\mathbf{s}$. A broad class of estimators for $\{\alpha, \beta\}$ are the solution to the optimisation problem:

$$\min_{\alpha,\beta} L(\mathbf{e}), \tag{5}$$

for some loss function L.

- $L(\mathbf{e}) = ||\mathbf{e}||_2 \rightarrow$ Least Squares (LS)
- $L(\mathbf{e}) = ||\mathbf{e}||_1 \rightarrow$ Least Absolute Deviations (LAD)

Question: It is well know that LS loses its desirable properties in the presence of heteroskedasticity. So, given daily financial return data, should we prefer LS or LAD?

Simulating data

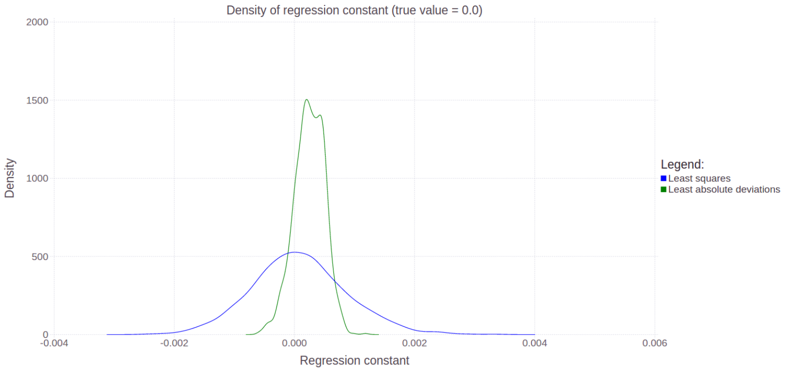
We want to simulate data using $r_t = \alpha + \beta s_t + e_t$.

- Let $v_{t,\kappa}$ denote moving window historical variance on NAB returns, where κ is window length
- Use $e_t \backsim \mathcal{N}(0, v_{t,\kappa})$ to simulate residuals
- Use $s_t \backsim \mathcal{N}(0, 1)$ to simulate signal
- Set $\alpha = 0$, and choose β such that $R^2 = 0.05$
- Simulate r_t

Note:

- Both r_t and e_t will exhibit the same pattern of heteroskedasticity (similar to NAB)
- $\kappa = 17$ results in robust kurtosis ≈ 4

Kernel density estimates for constant



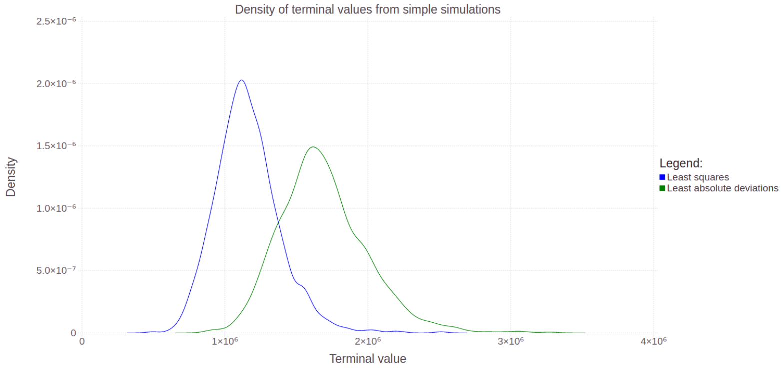
Kernel density estimates for coefficient

Implications for portfolio manager

It immediately follows that LAD yields more accurate predictions.
Question: Does this translate to better returns?

- Simple economy with 20 assets plus zero-interest cash asset
- Simulate all returns using method from previous slide
- Estimate $\{\alpha, \beta\}$ via LS and LAD using first half of sample
- For second half, hold (equal-weighted) in period t all assets with $\hat{r}_t = \hat{\alpha} + \hat{\beta} s_t > 0$

Kernel density estimates for terminal portfolio value



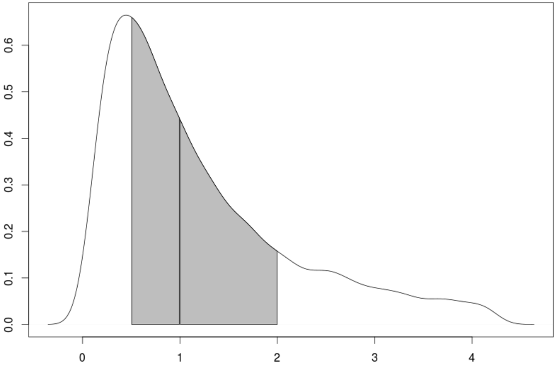
Portfolio terminal value for LS versus LAD estimation. Portfolio start value = 1 million.

Robust skew

Let Q_p denote the quantile corresponding to probability p. Then a robust measure of skewness is:

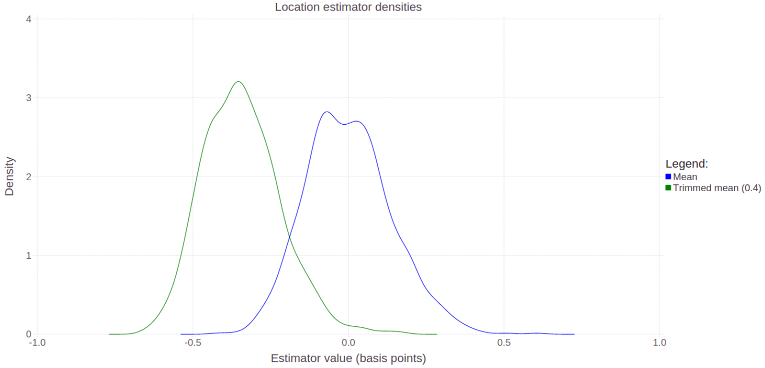
$$RobustSkewness = \frac{Q_{0.75} - Q_{0.5}}{Q_{0.75} - Q_{0.25}} - \frac{Q_{0.5} - Q_{0.25}}{Q_{0.75} - Q_{0.25}} \qquad (6)$$

Robust skew



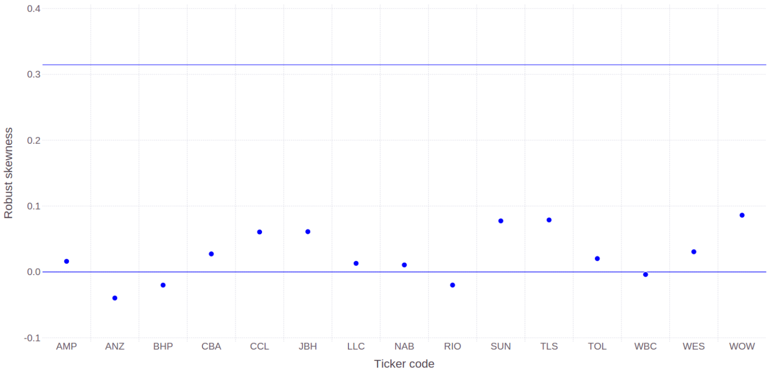
Kernel density of the Lognormal(0, 1) distribution with 0.25, 0.5, and 0.75 quantiles marked

Trim mean for Lognormal distribution



True mean $= 0$

Tail Fatness in Daily Financial Returns

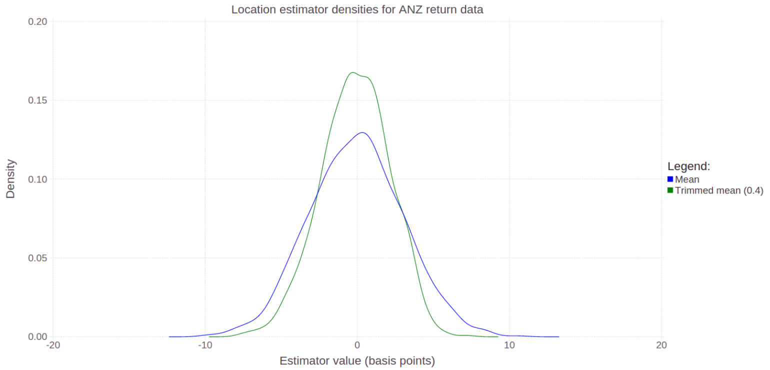


Robust skewness of daily financial returns for some popular stocks. Note, Symmetric and Lognormal(0, 1) lines included for reference.

Kernel density estimates for trimmed and sample mean

Resampled WOW daily return data (2005 to 2015).

Kernel density estimates for trimmed and sample mean



Resampled ANZ daily return data (2005 to 2015).

TL;DR

TL;DR: Any empiricist working with financial data subject to typical heteroskedasticity patterns should at least consider robust statistics.

All presentation materials and source code publicly available at:

https://github.com/colintbowers/RobustStatsTutorial.jl