Statistics for Financial Engineering: Some Examples

David Ruppert

Cornell University

Nov 21, 2008

Outline

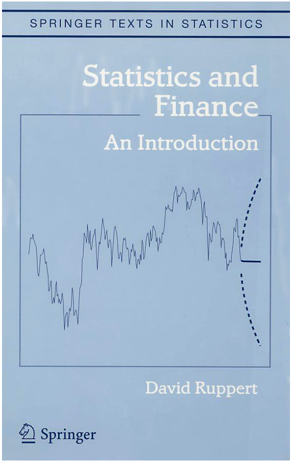
Some Themes

- Calibration of financial models is a statistical problem
- Many financial engineers come from math or physics and have little exposure to statistics and econometrics
- Financial engineering is an exciting area for statisticians to work in and to learn from

A little about myself

- BA and MA in mathematics
- PhD from in statistics in 1977
- taught in the statistics department at North Carolina for 10 years
- have been in Operations Research and Information (formerly Industrial) Engineering at Cornell since 1987
- starting teaching Statistics and Finance to undergraduates in 2001
  - textbook published in 2004
- starting teaching Statistics for Financial Engineering to master's students in 2008
  - working on revised and expanded textbook

Undergraduate Textbook



A little about my research

- have done research in
  - asymptotic theory of splines
  - semiparametric modeling
  - measurement error in regression
  - smoothing (nonparametric regression and density estimation)
  - transformation and weighting
  - stochastic approximation
  - biostatistics
  - environmental engineering
  - modeling of term structure
  - executive compensation and accounting fraud

Overview

- Recent example where a statistician could have helped
- Penalized splines
  - Two examples:
    - Return to interest rate dynamics
    - Term structure – estimating the forward rate curve
- Predicting the risk of accounting fraud

Three types of regression

Linear regression

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i, \ \ i = 1, \ldots, n$$

Nonlinear regression

$$Y_i = m(X_{i,1}, \ldots, X_{i,p}; \beta_1, \ldots, \beta_q) + \epsilon_i, \ \ i = 1, \ldots, n$$

where m is a known function depending on unknown parameters

Nonparametric regression

$$Y_i = m(X_{i,1}, \ldots, X_{i,p}) + \epsilon_i, \ \ i = 1, \ldots, n$$

where m is an unknown "smooth" function

Usual assumptions on the noise

Usually $\epsilon_1, \ldots, \epsilon_n$ are assumed to be:

- mutually independent (or at least uncorrelated)
- homoscedastic (constant variance)
- normally distributed

Much research over the last 50+ years has looked into ways of

1. checking these assumptions
2. statistical methods that require less assumptions

Estimation of Default Probabilities

Data:

- ratings: 1=Aaa (best),...,16=B3 (worse)
- default frequency (estimate of default probability)

Some statistical models

- **nonlinear model:**

$$\Pr(\text{default}|\text{rating}) = \exp\{\beta_0 + \beta_1 \text{rating}\}$$

- **linear/transformation model (in recent textbook):**

$$\log\{\Pr(\text{default}|\text{rating})\} = \beta_0 + \beta_1 \text{rating}$$

  - **Problem:** cannot take logs of default frequencies that are 0
  - **(Sub-optimal) solution in textbook:** throw out these observations

A better statistical model

- **Transform-both-sides (TBS) model** – see Carroll and Ruppert (1984, 1988):
- using a power transformation:

$$\Pr(\text{default}|\text{rating})^{\alpha} = \exp[\alpha\{\beta_0 + \beta_1 \text{rating}\}]$$

  - α chosen by residual plots (or maximum likelihood)
  - $\alpha = 1/2$ works well for this example
  - log transformations are also commonly used

The Box-Cox family

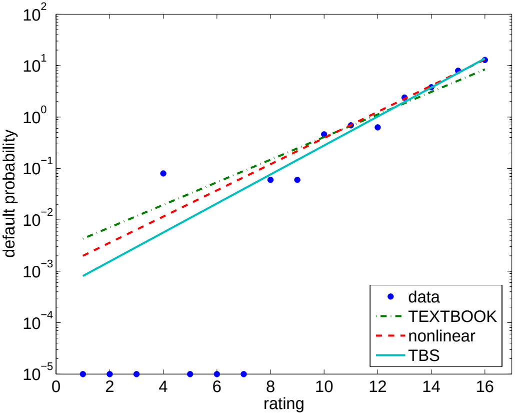
- the most common transformation family is due to Box and Cox (1964):

$$\begin{aligned} h(y,\lambda) &= \frac{y^\lambda - 1}{\lambda} \text{ if } \lambda \neq 0 \\ &= \log(y) \text{ if } \lambda = 0 \end{aligned}$$

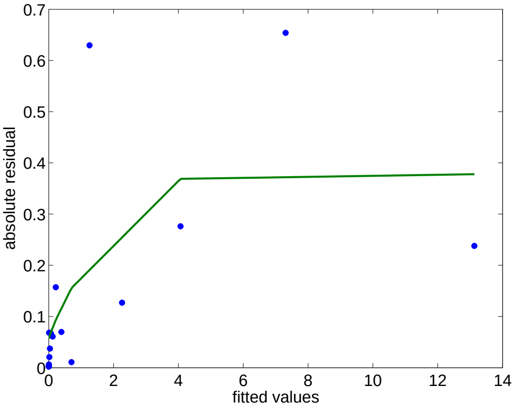
- derivative has simple form:

$$h_y(y,\lambda) = \frac{d}{dy} h(y,\lambda) = y^{\lambda-1} \text{ for all } \lambda$$

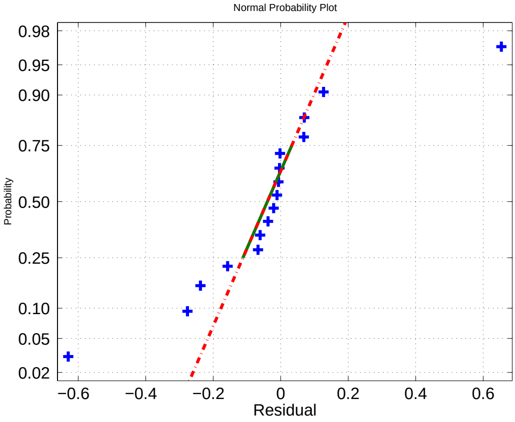
TBS fit compared to others



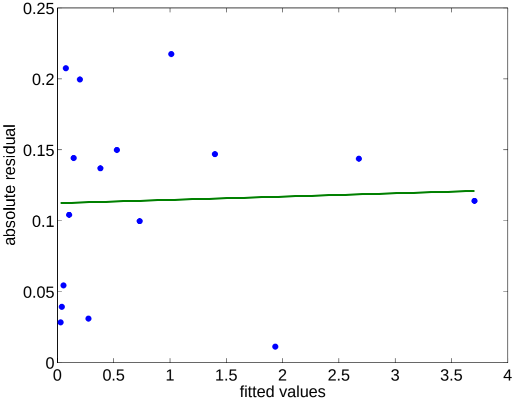
Nonlinear regression residuals



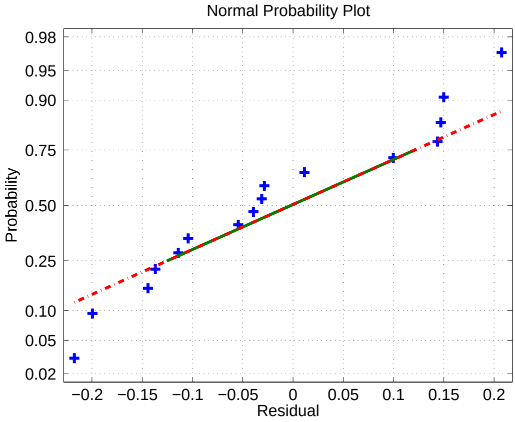
Nonlinear regression residuals



TBS residuals



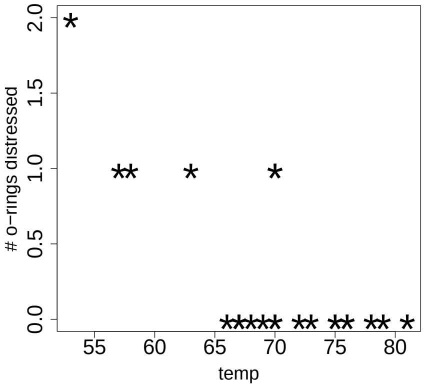
TBS residuals



Estimated default probabilities

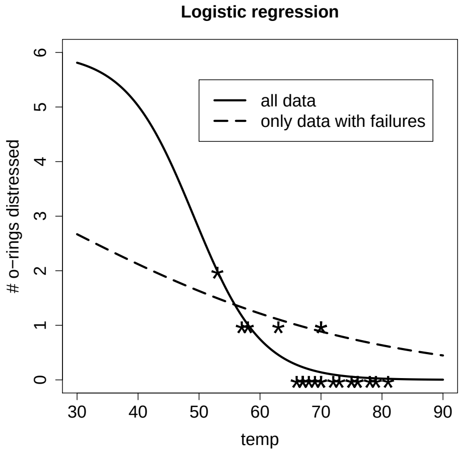
method	$\widehat{Pr}\{\text{default}\|\text{Aaa}\}$	as % of TEXTBOOK est
TEXTBOOK	0.005%	100%
nonlinear	0.002%	40%
TBS	0.0008%	16%

A Similar Problem: Challenger Data

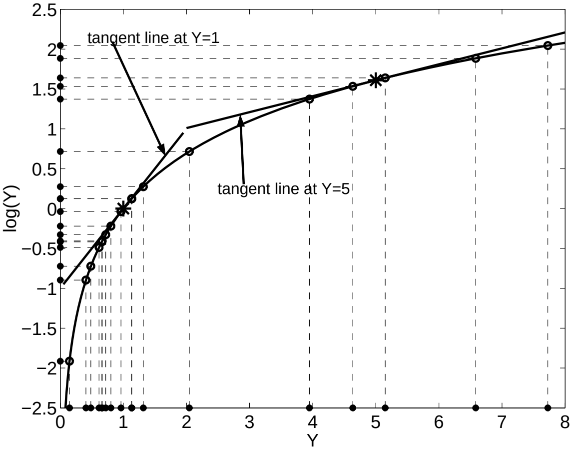


Data from UCI Machine Learning Repository (Donor: David Draper)

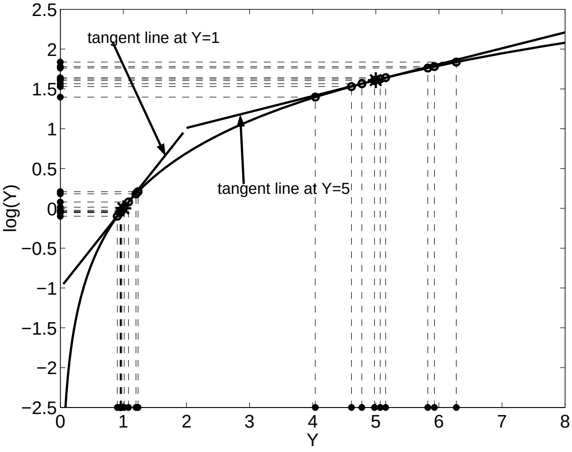
Challenger Data: Extrapolation to 31°



Normalizing transformation: how it works



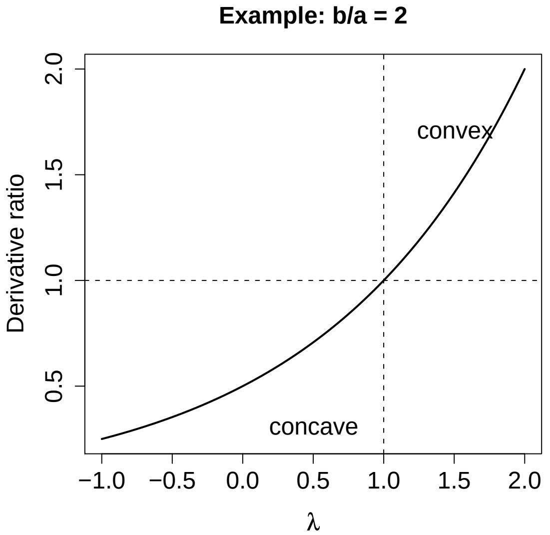
Variance stabilizing transformation: how it works



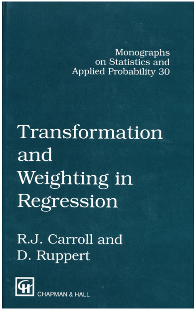
Strength of Box-Cox family

- Take $a < b$
- Then
$$\frac{h_y(b,\lambda)}{h_y(a,\lambda)} = \left(\frac{b}{a}\right)^{\lambda-1}$$
which is increasing in λ and equals 1 when $\lambda = 1$
- $\lambda = 1$ is the dividing point between concave and convex transformations
- $h(y,\lambda)$ becomes a stronger concave transformation as λ decreases from 1
- also, $h(y,\lambda)$ becomes a stronger convex transformation as λ increases from 1

Strength of Box-Cox family, cont.



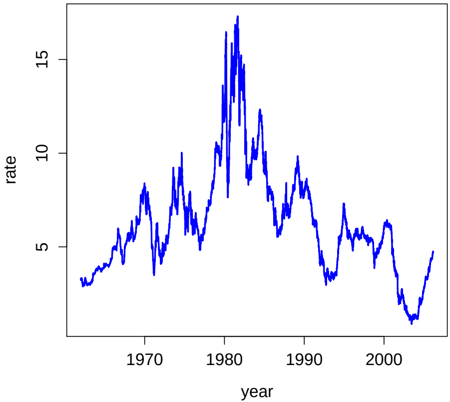
Reference for TBS



Transformation and Weighting in Regression by Carroll and Wand (1988)

- Lots of examples
- But none in finance ☹

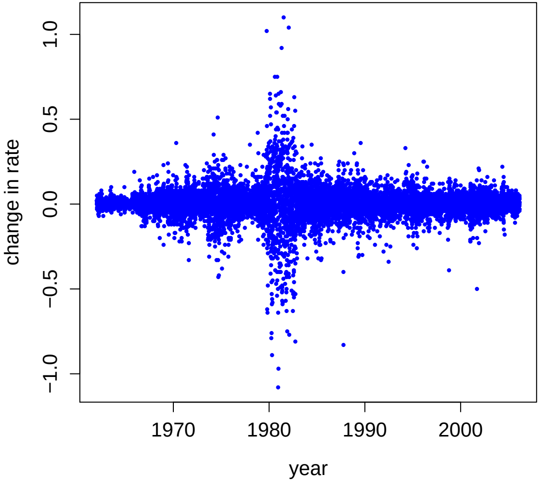
1-Year Treasury Constant Maturity Rate, daily data



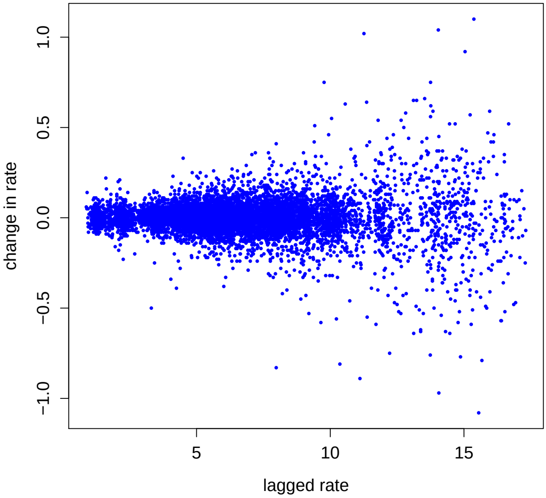
Source: Board of Governors of the Federal Reserve System

http://research.stlouisfed.org/fred2/

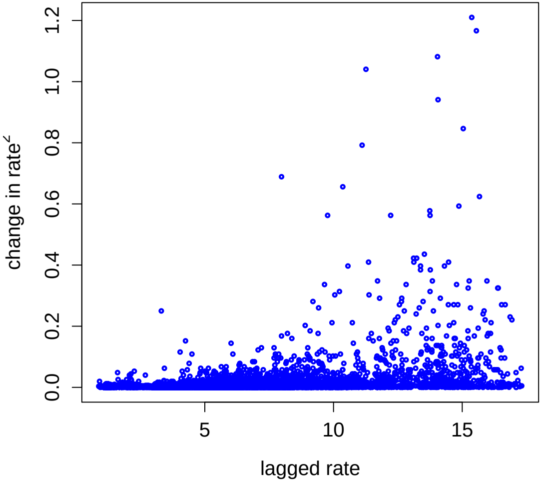
ΔR_t versus year



ΔR_t versus R_{t-1}



ΔR_t^2 versus R_{t-1}



Drift function

Discretized diffusion model:

$$\Delta R_t = \mu(R_{t-1}) + \sigma(R_{t-1})\epsilon_t$$

- $\mu(x)$ is the drift function
- $\sigma(x)$ is the volatility function (as before)

Estimating Volatility

Parametric model:

$$\mathrm{Var}\{(\Delta R_t)\} = \beta_0 R_{t-1}^{\beta_1}$$

(Common in practice)

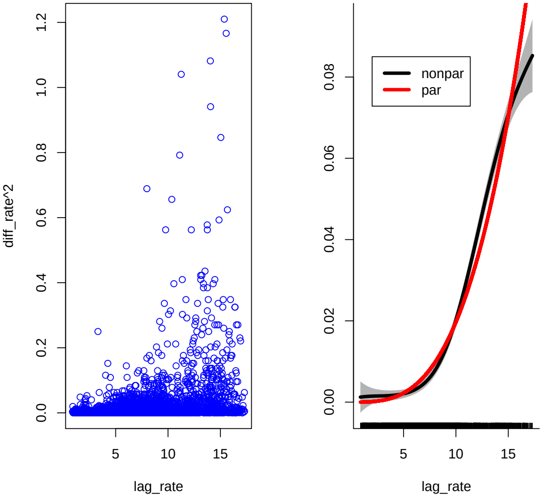
Nonparametric model:

$$\mathrm{Var}\{(\Delta R_t)\} = \sigma^2(R_{t-1})$$

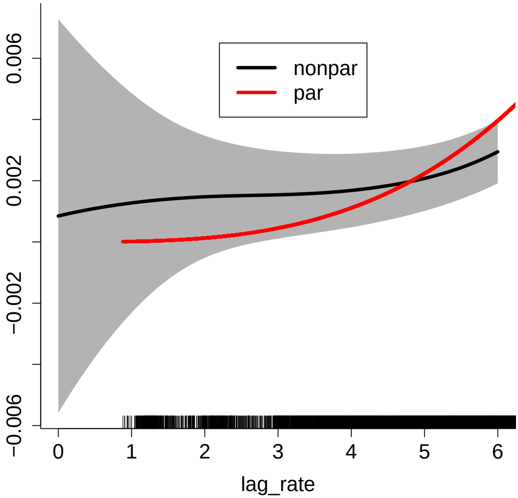
where $\sigma(\cdot)$ is a smooth function

- will be modeled as a spline
- In these models: no dependence on t

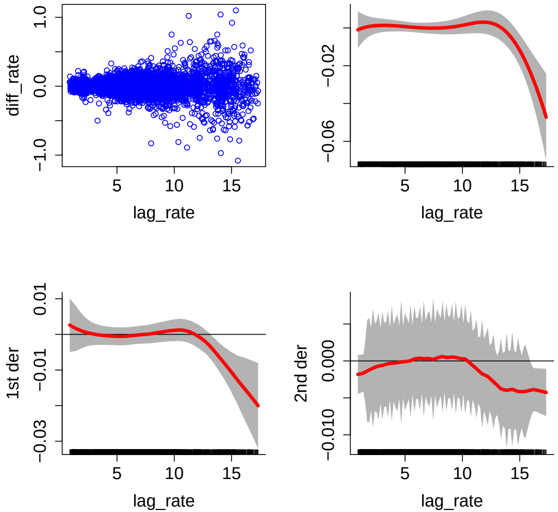
Comparing parametric and nonparametric volatility fits



Comparing parametric and nonparametric volatility fits: zooming in near 0



Spline fitting – Estimation of drift function



Splines

Nonparametric regression model:

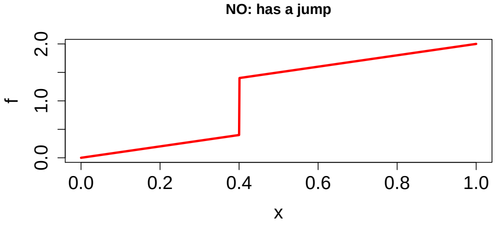
$$Y_i = m(X_i) + \epsilon_i.$$

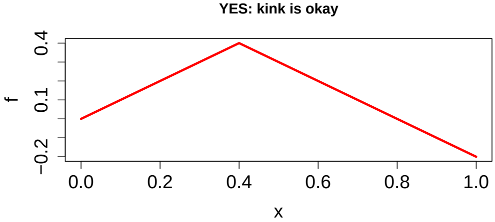
Here $m(\cdot)$ is a smooth function. We will model it as a spline.

- p = degree ($p + 1$ = order)
- $\kappa_0, \kappa_1 < \cdots < \kappa_K < \kappa_{K+1}$ = the knots
- $m(x)$ is a polynomial of degree p on $\kappa_{k-1} < x < \kappa_k$ for $k = 1, \ldots, K+1$
- $m^{p-1}(x)$ is continuous everywhere
  - but $m^p(x)$ can jump at the knots
  - so splines have "maximal smoothness"

Splines have "maximal smoothness"

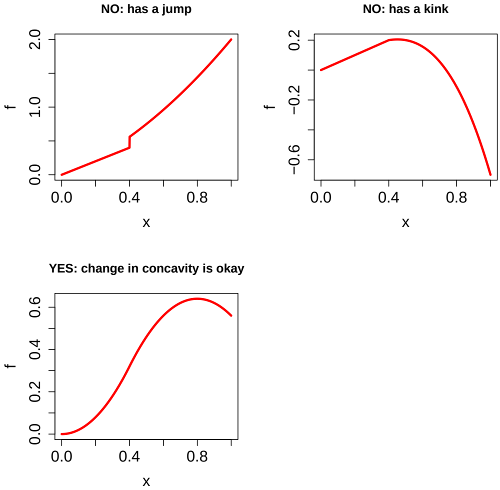
Case 1: $(p = 1)$ Is this a linear spline?





Splines have "maximal smoothness," cont.

Case 2: $(p = 2)$ Is this a quadratic spline?



Residuals for diffusion model

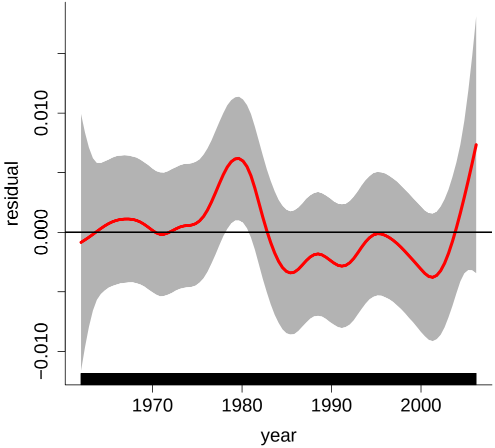
$$\begin{aligned}\text{residual}_t &:= \Delta R_t - \widehat{\mu}(R_{t-1}) \\ E(\text{residual}_t) &= 0\end{aligned}$$

$$\begin{aligned}\text{std residual}_t &:= \frac{\text{residual}_t}{\widehat{\sigma}(R_{t-1})} \\ E(\text{std residual}_t^2) &= 1\end{aligned}$$

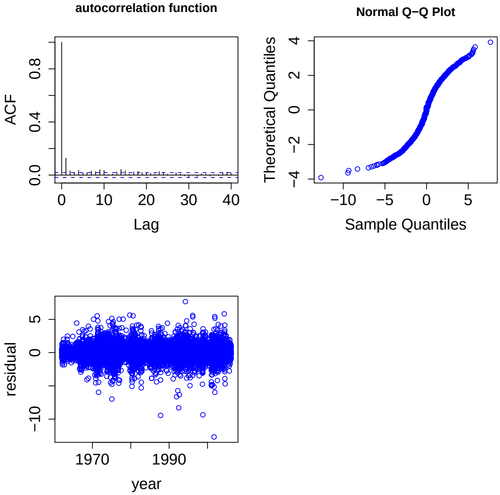
Question

Are the drift and volatility functions constant in time?

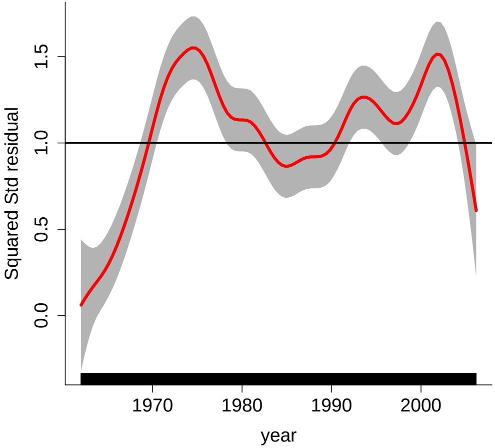
Residual plots: ordinary residuals



Residual plots: standardized residuals



Residual plots: Squared standardized residuals



GARCH(p, q) model

The GARCH(p, q) model is

$$a_t = \epsilon_t \sigma_t,$$

where

$$\sigma_t = \sqrt{\alpha_0 + \sum_{i=1}^{q} \alpha_i a_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2}.$$

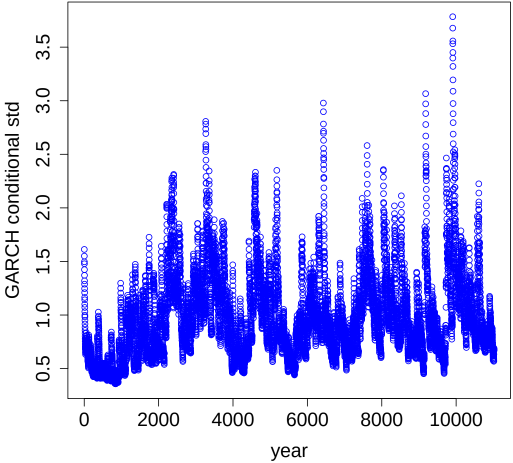
and

$$\epsilon_t \text{ is an independent white noise process}$$

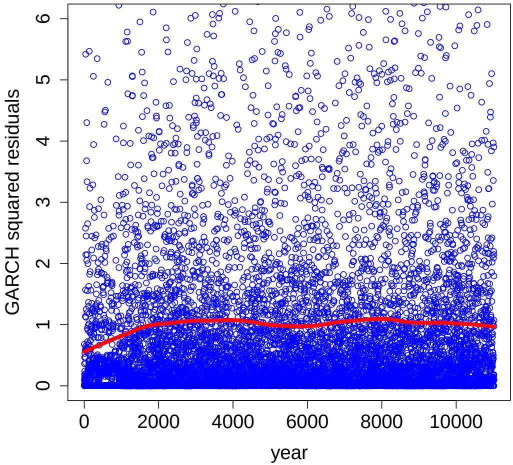
GARCH(1,1) fit


```
Call:
garch(x = std_drift_resid^2, order = c(1, 1))

Model:
GARCH(1,1)

Residuals:
     Min        1Q    Median        3Q       Max
7.66e-11 1.79e-02 8.73e-02 3.35e-01 2.49e+01

Coefficient(s):
    Estimate  Std. Error  t value Pr(>|t|)
a0   0.27291     0.00148      184   <2e-16 ***
a1   0.44690     0.00252      177   <2e-16 ***
b1   0.80490     0.00075     1073   <2e-16 ***

        Box-Ljung test

data:  Squared.Residuals
X-squared = 0.13, df = 1, p-value = 0.7186
```

GARCH: estimated conditional standard deviations



GARCH: squared residuals with lowess smooth

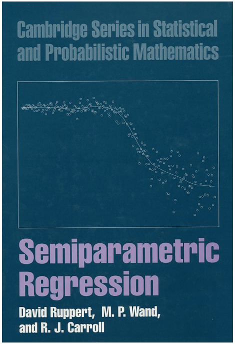


Final model for the interest rate dynamics

$$\Delta R_t = \mu(R_{t-1}) + \sigma(R_{t-1})\sigma_{\mathrm{GARCH}}(t)\,\epsilon_t$$

1. Model was fit in two steps:
   1. estimate $\mu()$ and $\sigma()$
   2. estimate $\sigma_{\mathrm{GARCH}}(t)$
2. Could the two step be combined?
3. Would combining them change the results?

Reference for spline modeling



Semiparametric Regression by Ruppert, Wand, and Carroll (2003)

- Lots of examples.
- But most from biostatistics and epidemiology

Software

The spline fits shown here were obtained using the function `spm`

- in `R`'s `SemiPar` package
- author is Matt Wand

Estimating Term Structure with a Spline Model

Joint work with:

- Bob Jarrow (Cornell)
- Yan Yu (University of Cincinnati)

Bond prices and the forward rate

- t = time to maturity
- $f(t)$ = forward rate
- $F(t) = \int_0^t f(s)ds$
- $y(t)$ = yield to maturity $= F(t)/t$
- $P(t)$ = price of zero-coupon bond
- $D(t)$ = discount function

$$\begin{aligned}\frac{P(t)}{\text{PAR}} &= D(t) = \exp\{-F(t)\} \\ &= \exp\{-ty(t)\} = \exp\left\{-\int_0^t f(s)ds\right\}.\end{aligned}$$

Estimation of the forward rate

Suppose the ith bond pays $C_i(t_{i,j})$ and time $t_{i,j}$

- $i = 1, \ldots, n$
- $j = 1, \ldots, z_i$

Let $f(s, \boldsymbol{\delta}) = \boldsymbol{\delta}' \boldsymbol{B}(s)$ be a spline model for the forward rate.

- $\boldsymbol{B}(s)$ is a vector of spline basis functions (many choices)

Model for price of ith bond:

$$\widehat{P_i}(\boldsymbol{\delta}) = \sum_{j=1}^{z_i} C_i(t_{i,j}) \exp\{-\boldsymbol{\delta}' \boldsymbol{B}^I(t_{i,j})\}$$

where

$$\boldsymbol{B}^I(t) = \int_0^t \boldsymbol{B}(s) ds$$

Corporate Bonds

- Problem: often there are not enough bonds to fit a fully nonparametric model
- Jarrow, Ruppert, and Yu solve this by using a semiparametric model

Algorithm

Step 1: Nonparametric P-spline model for the forward rate to US Treasury bonds.

- $\boldsymbol{\delta}$ is estimated by penalized least squares
- $\widehat{f}_{Tr}(t) = \widehat{\boldsymbol{\delta}}' \boldsymbol{B}(t)$, where $\widehat{\boldsymbol{\delta}}$ are the estimated spline coefficients

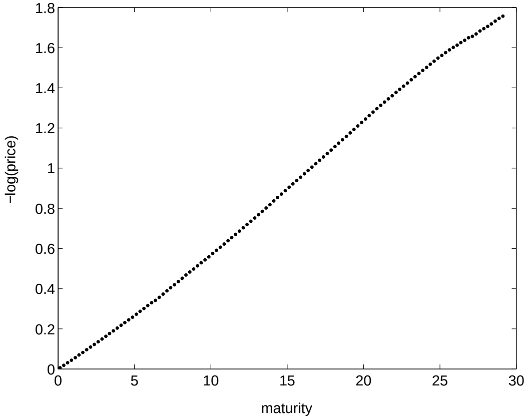
Step 2: Parametric estimation to obtain the forward rate curve for a corporation's bonds.

- credit spread is parametric with parameter α
- for example, if the credit spread is a constant, then

$$f_C(t) = \widehat{f}_{Tr}(t) + \alpha = \hat{\boldsymbol{\delta}}' \boldsymbol{B}(t),$$

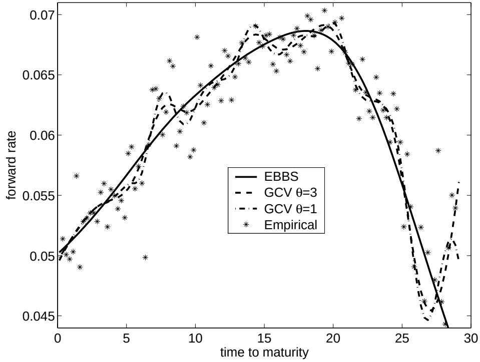
- fix $\hat{\boldsymbol{\delta}}$ at value from Step 1 and estimate α by OLS

−Log-prices (as fraction of PAR)



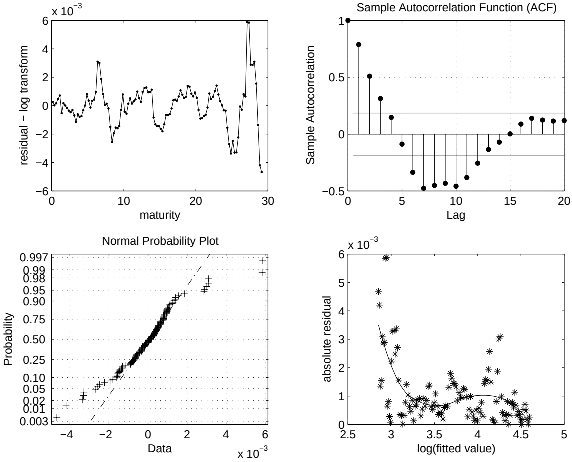
Data from Dec 1995

Estimates of forward rate

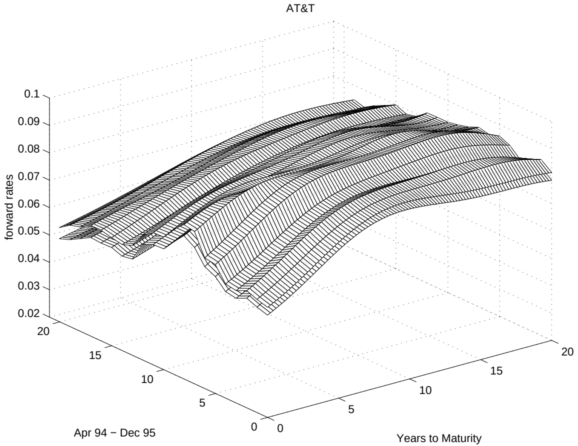


$$\text{empirical forward} = -\frac{\log\{P(t_{i+1})\} - \log\{P(t_i)\}}{t_{i+1} - t_i}$$

Residual analysis



Estimates of Treasury and AT&T forward rates



Reference on modeling term structure

Jarrow, R., Ruppert, D., and Yu, Yan. (2004) Estimating the term structure of corporate debt with a semiparametric penalized spline model, JASA, 99, 57–66.

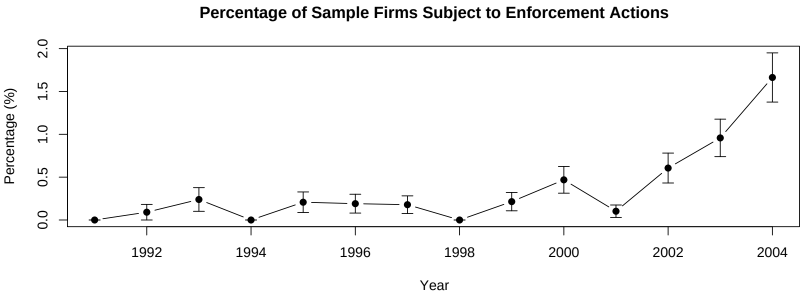
Predicting the risk of accounting fraud

- manuscript: "Predicting the risk of SEC enforcement using executive compensation data"
- authors:
  - Emmanuel Sharef – now at Morgan-Stanley
  - David Ruppert
- uses survival analysis
  - to handle censoring
- similar to applications used to predict bankrupcy (Chava and Jarrow, 2004)

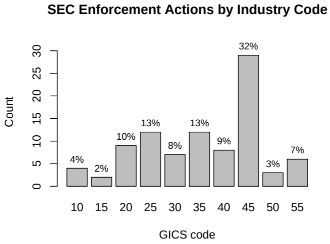
Data Sources

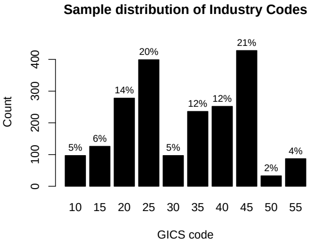
- AAER (Accounting and Auditing Enforcement Releases) from SEC
  - we used only firms traded on NYSE, AMEX, or Nasdaq
  - 1991 to 2004
  - 92 “failures” = AAER issued to firm
- executive compensation data from EXECUCOMP database
- accounting data from COMPUSTAT

Baseline enforcement probability



Enforcement actions by industry code





Ind. Code	Industry Name	Enforcement Count (%)	Sample Count (%)	Enforcement Frequency
10	Energy	4 (4.3%)	97 (4.8%)	4.1%
15	Materials	2 (2.2%)	126 (6.2%)	1.6%
20	Industrials	9 (9.8%)	278 (13.7%)	3.2%
25	Consumer Discretionary	12 (13%)	399 (19.6%)	3.0%
30	Consumer Staples	7 (7.6%)	97 (4.8%)	7.2%
35	Health Care	12 (13%)	236 (11.6%)	5.1%
40	Financials	8 (8.7%)	252 (12.4%)	3.2%
45	Information Technology	29 (31.5%)	428 (21.1%)	6.7%
50	Telecommunication	3 (3.3%)	33 (1.6%)	9.1%
55	Utilities	6 (6.5%)	87 (4.3%)	6.9%
Total		92 (100%)	2033 (100%)	4.5%

Notation for survival model

- $[T_0, T_1]$ = time-frame of study
- t_i = time when ith firm starts equity trading
- τ_i = time of enforcement action of ith firm
- C_i = censoring time of ith firm
- $Y_i = \min(\tau_i, C_i)$
- δ_i censoring indicator
- X_i = covariate vector (accounting and executive compensation variables)

Hazard Model

Discrete-time hazard rate:

$$P_t^i = \mathbb{P}\left\{\tau_i = t | \tau_i \geq t, X^i_{t_i + a_i}, \ldots, X^i_{\tau_i \wedge t}\right\}$$

Logistic regression model:

$$P_t^i = \frac{1}{1 + e^{-\beta^T X_t^i}}$$

Estimation by maximizing the survival analysis likelihood.

Model Selection And Goodness-Of-Fit Testing

- 1991–2001 = training data
- 2003–2004 = test data (not used for estimation or model selection)
- Model selection by cross-validation
  - Use 1991-1995 to predict 1996-1997
  - Use 1991-1996 to predict 1997-1998
  - and so forth
  - finally, use 1991-2000 to predict 2001-2002
  - Use ROC curves to check for goodness-of-fit

Most Significant Effects

	Estimate	Std Error	z value	p-value
sale3ls	−0.01579	0.00731	−2.16	0.03
bs_volatility	1.25500	0.34220	3.67	0.00
ceo_othann_log_25_d	0.66180	0.33470	1.98	0.05
ceo_options_log_25_d	0.74750	0.34180	2.19	0.03
shrownpc	0.05365	0.02130	2.52	0.01
ceo_is_chman	1.26800	0.48600	2.61	0.01
auditor_top5_d	−0.79080	0.32900	−2.40	0.02

sale3ls 3-year least squares annual growth of sales

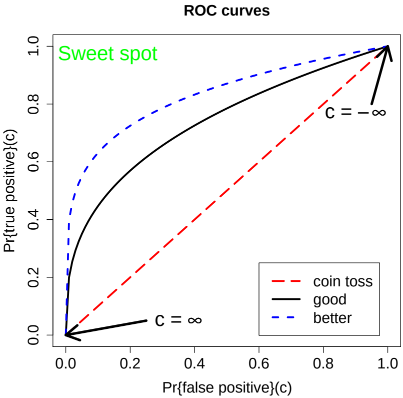
bs_volatility Black-Scholes volatility over 60 months

ceo_othann_log_25_d = 1 if CEO's other annual compensation in top 25th percentile

ceo_options_log_25_d = 1 if value of stock options to CEO during the year in the top 25th percentile

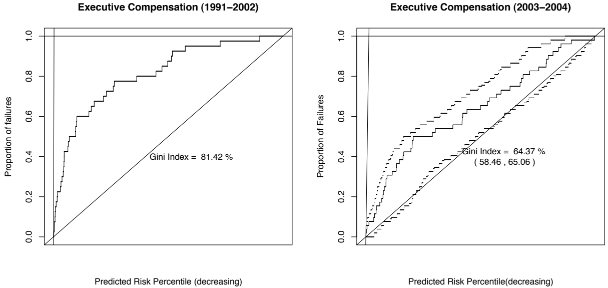
shrownpc percentage of company's shares owned by CEO.

ROC curve: what is it?

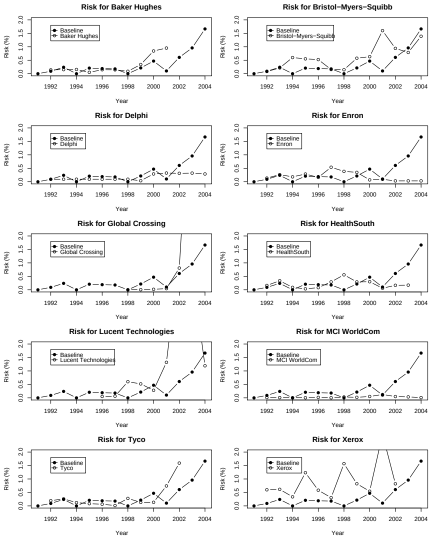


- "signal" is used to detect positives
- c is a "tuning parameter"
- signal $> c \Rightarrow$ positive

ROC curves


Executive Compensation (1991–2002)
Proportion of failures
0.0
0.2
0.4
0.6
0.8
1.0
Gini Index = 81.42 %
Predicted Risk Percentile (decreasing)
Executive Compensation (2003–2004)
Proportion of Failures
0.0
0.2
0.4
0.6
0.8
1.0
Gini Index = 64.37 %
(58.46 , 65.06)
Predicted Risk Percentile(decreasing)

Examples


Risk for Baker Hughes
Baseline
Baker Hughes
Risk for Bristol−Myers−Squibb
Baseline
Bristol−Myers−Squibb
Risk for Delphi
Baseline
Delphi
Risk for Enron
Baseline
Enron
Risk for Global Crossing
Baseline
Global Crossing
Risk for HealthSouth
Baseline
HealthSouth
Risk for Lucent Technologies
Baseline
Lucent Technologies
Risk for MCI WorldCom
Baseline
MCI WorldCom
Risk for Tyco
Baseline
Tyco
Risk for Xerox
Baseline
Xerox
Risk (%)
Year
0.0
0.5
1.0
1.5
2.0
1992
1994
1996
1998
2000
2002
2004

Thanks

Thank you for coming