Linear Regression in Finance

FA590 Statistical Learning in Finance

Dr. Zonghao Yang

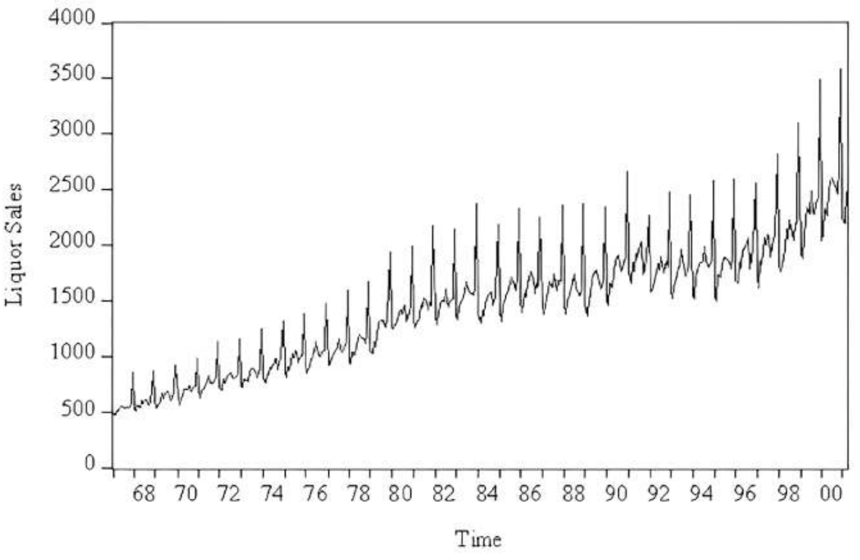
Stevens Institute of Technology

Part I

Trend and Seasonality of Financial Time Series

Trend and Seasonality in Financial Time Series

Question How do you model trend and seasonality?



Modeling Trend

- Is the series continuously increasing or decreasing?
- For many statistical analysis, especially time series analysis, we require stationarity: Means and variances do not change over time
- If a series continuously grows or declines, its mean is changing – not stationary

Modeling Trend

- Long-term movement of a series

$$T_t = \beta_0 + \beta_1 \text{TIME}_t$$

- Quadratic trend

$$T_t = \beta_0 + \beta_1 \text{TIME}_t + \beta_2 \text{TIME}_t^2$$

- Exponential trend

$$T_t = \beta_0 e^{\beta_1 \text{TIME}_t}$$

$$\log(T_t) = \log(\beta_0) + \beta_1 \text{TIME}_t$$

- TIME_t is a deterministic variable, e.g. 1, 2, ..., T

Modeling Trend

We can estimate trend models using the same techniques as for OLS

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0,\beta_1} \sum_{t=1}^{T} (y_t - \beta_0 - \beta_1 \mathrm{TIME}_t)^2$$

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \arg\min_{\beta_0,\beta_1,\beta_2} \sum_{t=1}^{T} \left(y_t - \beta_0 - \beta_1 \mathrm{TIME}_t - \beta_2 \mathrm{TIME}_t^2\right)^2$$

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0,\beta_1} \sum_{t=1}^{T} \left(y_t - \beta_0 e^{\beta_1 \mathrm{TIME}_t}\right)^2$$

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0,\beta_1} \sum_{t=1}^{T} (\log(y_t) - \log(\beta_0) - \beta_1 \mathrm{TIME}_t)^2$$

FORECASTING TREND

- Forecast

$$\hat{y}_{T+h,T} = \hat{\beta}_0 + \hat{\beta}_1 \text{TIME}_{T+h}$$

- Interval forecast

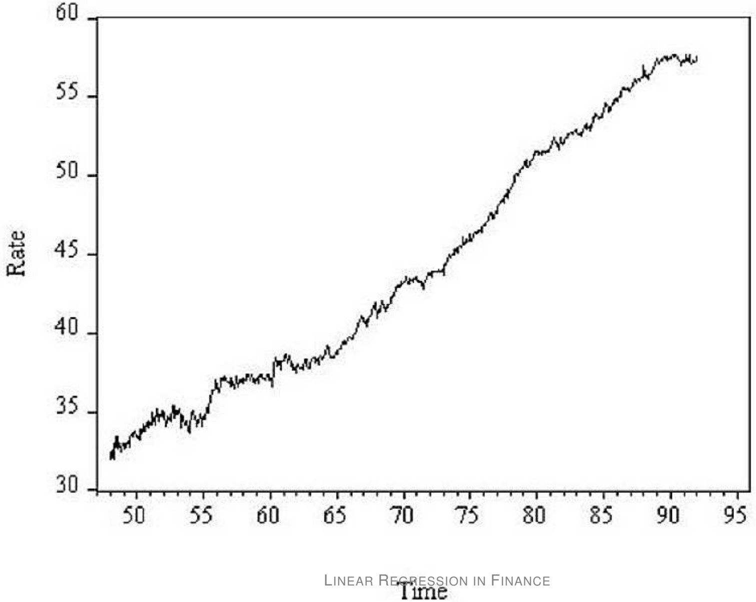
$$[\hat{y}_{T+h,\tau} - 2\hat{\sigma}, \hat{y}_{T+h,\tau} + 2\hat{\sigma}]$$

- Density forecast:

$$N(\hat{y}_{T+h,\tau}, \hat{\sigma}^2)$$

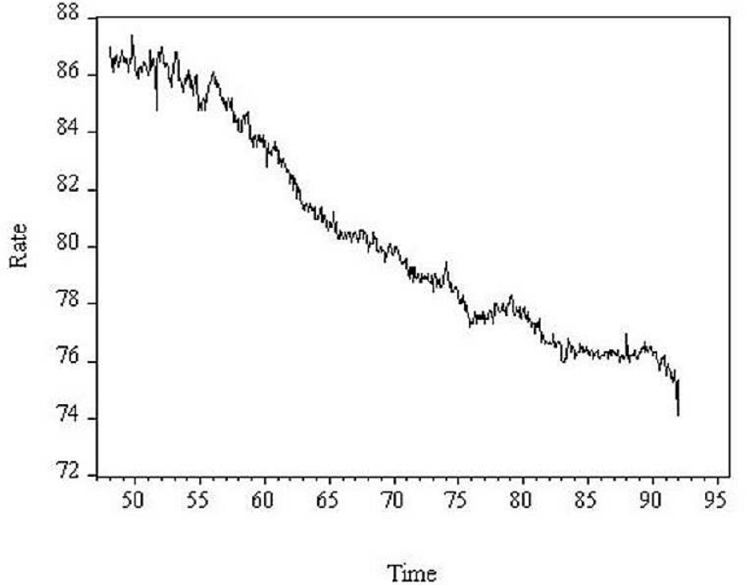
Example: Labor Force Participation

- U.S. Female labor force participation rate over time
  - What do we see from this time series plot?
  - How should we model this series?



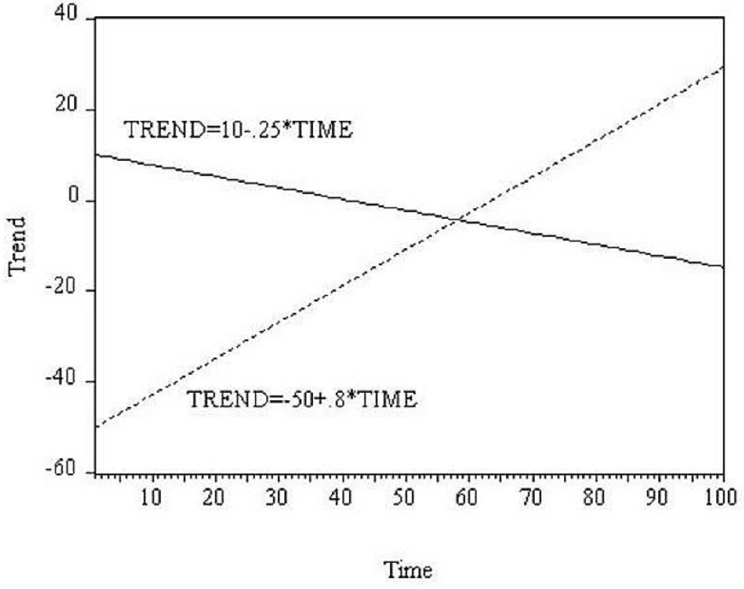
Example: Labor Force Participation

- U.S. Male labor force participation rate over time
  - How do we model this series?



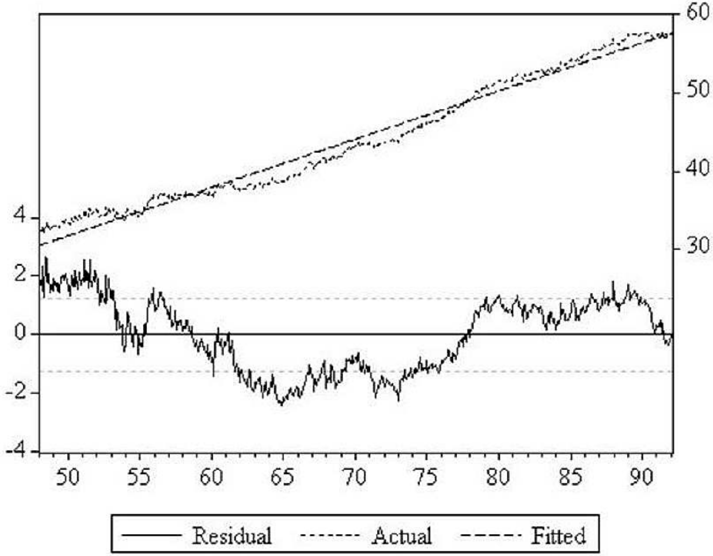
Example: Labor Force Participation

- Two trend lines



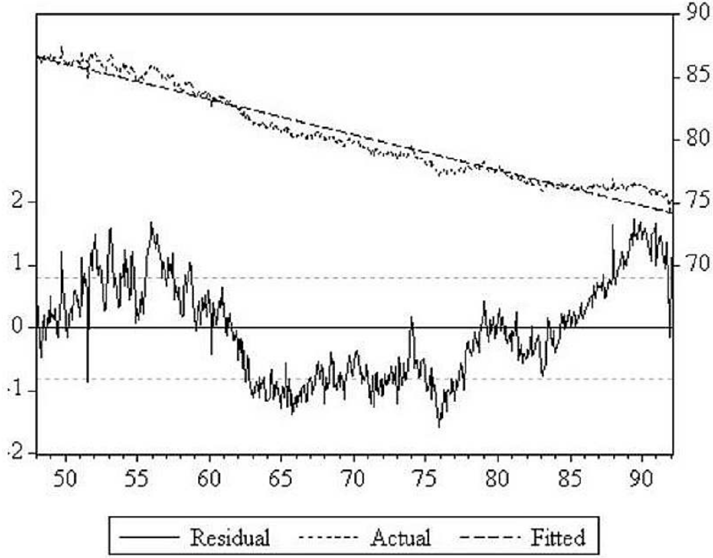
Example: Labor Force Participation

- Female labor force with trend line, note the two vertical axes



Example: Labor Force Participation

- Male labor force with trend line

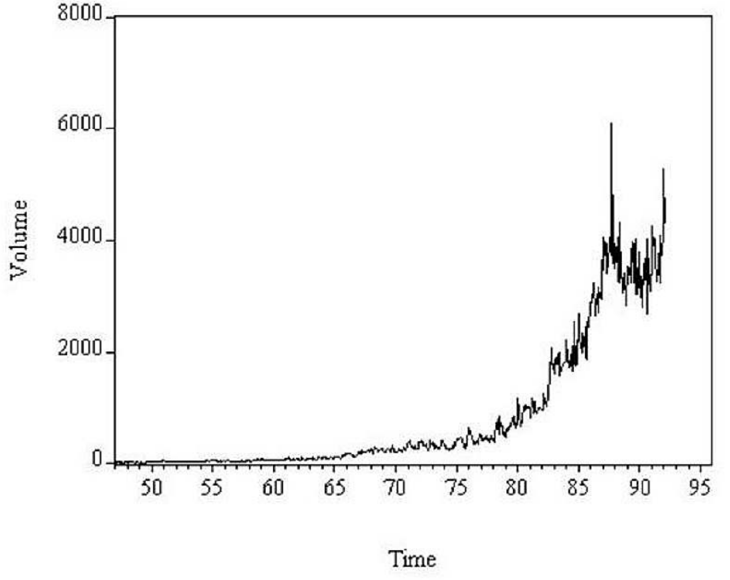


Example: Labor Force Participation

- Female and male raw series look totally different
- After removing the trend, they don't look so different!
- Time trend can obscure more subtle variation
- Model trend first, then look for more detailed variation

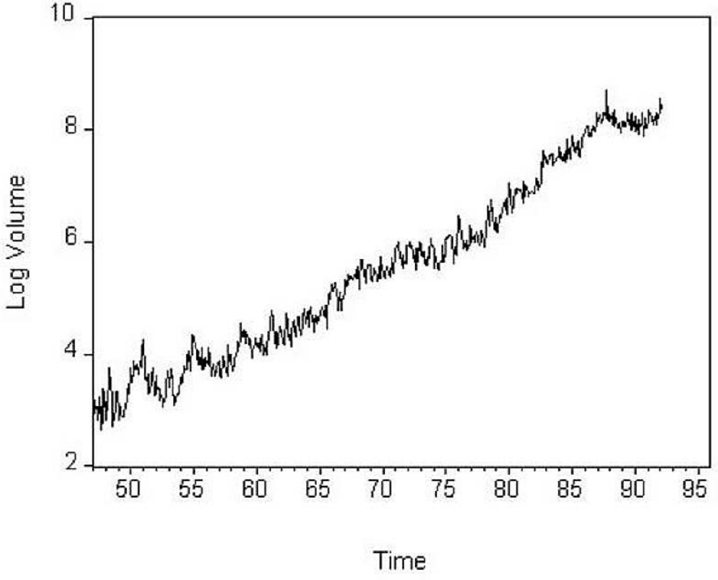
Example: Trading Volume

- Trading volume on the New York Stock Exchange
  - How would we model this series?



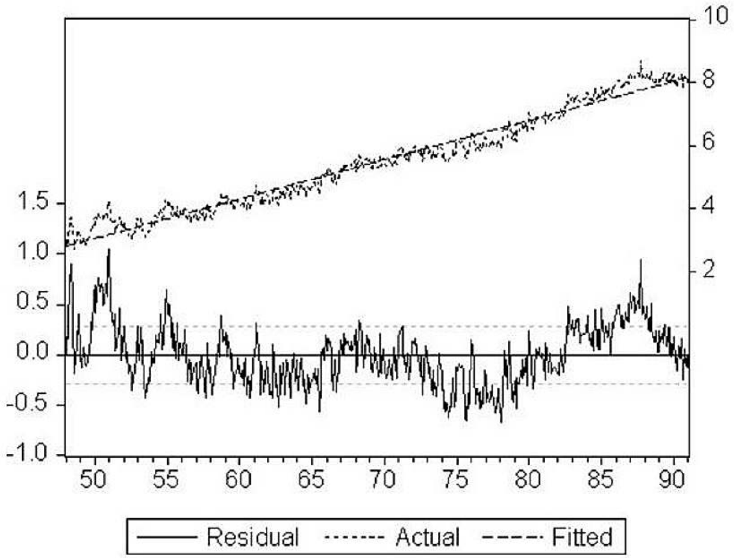
Example: Trading Volume

- Log volume on the NYSE: Looks linear!



Example: Trading Volume

- Exponential trend on NYSE volume data



Modeling Seasonality

- **Seasonality** Periodic changes to the data
- The changes are predictable
- Example: Beaches are more crowded in the summer than the winter
- Example: Gasoline sales; People drive more in the summer
- Example: Traffic is worse from 4 pm – 7 pm than 12 pm – 3 pm
  - But not on holidays: Interaction between two seasonal trends

Modeling Seasonality

- We use indicator variables at regular intervals to capture seasonality
- For example, indicator variables for each quarter

$$D_{1,t} = \begin{cases} 1 & \text{if time } t \text{ is Spring} \\ 0 & \text{otherwise} \end{cases}$$

$$D_{2,t} = \begin{cases} 1 & \text{if time } t \text{ is Summer} \\ 0 & \text{otherwise} \end{cases}$$

$$D_{3,t} = \begin{cases} 1 & \text{if time } t \text{ is Autumn} \\ 0 & \text{otherwise} \end{cases}$$

$$D_{4,t} = \begin{cases} 1 & \text{if time } t \text{ is Winter} \\ 0 & \text{otherwise} \end{cases}$$

Modeling Seasonality

- We model seasonality by incorporating the indicator variables in our model

$$S_t = \sum_{i=1}^{s} \gamma_i D_{i,t}$$

- If all the γ's are the same, we do not have seasonality, and we can drop S_t
- Could have more than one type of seasonality

$$S_t = \sum_{i=1}^{s} \gamma_i D^1_{i,t} + \sum_{i=1}^{v} \delta_i D^2_{i,t}$$

Forecasting Seasonality

- Realizations

$$y_t = S_t + \epsilon_t$$

$$y_t = \sum_{i=1}^{s} \gamma_i D_{i,t} + \epsilon_t$$

- Forecast

$$\hat{y}_{T+h} = \sum_{i=1}^{s} \hat{\gamma}_i D_{i,T+h}$$

- Interval forecast

$$[\hat{y}_{T+h} - 2\hat{\sigma}, \hat{y}_{T+h} + 2\hat{\sigma}]$$

- Density forecast

$$N(\hat{y}_{T+h}, \hat{\sigma}^2)$$

Model Trend and Seasonality

- Could combine trend and seasonality

$$y_t = \beta_1 TIME_t + \sum_{i=1}^{s} \gamma_i D_{i,t} + \epsilon_t$$

$$y_t = \beta_1 TIME_t + \sum_{i=1}^{s} \gamma_i D^1_{i,t} + \sum_{i=1}^{v} \delta_i D^2_{i,t} + \epsilon_t$$

Part II

Regression in Asset Pricing

Types of Economic Data

- **Cross section**
  - Notation: $i = 1, 2, \ldots, N$
- **Time series**
  - Notation: $t = 1, 2, \ldots, T$
- **Panel data** Having both cross-section and time-series dimensions

permno	DATE	mvel1	beta	betasq	chmom	dolvol	idiovol	indmom	mom1m
12416	1985-11-29	37398.000000	0.102588	0.010524	-0.139829	9.179391	0.022234	0.188869	-0.007194
80038	1994-08-31	16150.000000	NaN	NaN	NaN	8.977494	NaN	-0.020070	0.134328
66472	2001-02-28	10313.125000	0.618998	0.383159	0.155628	8.379826	0.105263	-0.223246	-0.070513
85005	2004-07-30	44004.478386	1.153467	1.330486	-0.849747	11.449005	0.111983	0.567376	-0.258929
92523	2011-02-28	278257.595650	0.510935	0.261055	-0.009013	10.594065	0.028881	0.088227	0.034775

Figure. Panel Data from Final Project

Types of Regression in Asset Pricing

- **Time-series regression**
  - Analyzes a single asset over multiple time periods
  - The goal is to see how sensitive the asset's return is to various systematic risk factors (e.g., market risk premium, size premium,value premium) over time
- **Cross-sectional regression**
  - Analyzes many assets at a single point in time
  - The goal is to see which firm characteristics (e.g., size, book-to-market ratio, momentum) explain the differences in returns across those assets during that period
- **Panel regression**
  - Combine both the dimensions, analyzing multiple assets over multiple time periods
  - The goal is to provide a more powerful test by controlling for both time-invariant asset characteristics and market-wide time effects

Time-Series Regression

The general form for a single asset i is:

$$R_{i,t} - R_{f,t} = \alpha_i + \sum_{k=1}^{K} \beta_{i,k} F_{k,t} + \epsilon_{i,t} \quad \text{for } t = 1, \ldots, T$$

- $R_{i,t} - R_{f,t}$: The excess return of asset i at time t
- $F_{k,t}$: The realization of the k-th systematic risk factor at time t
- $\beta_{i,k}$: The **factor loading** or sensitivity of asset i to risk factor k
- α_i: **Jensen's Alpha**, representing the abnormal return of the asset after accounting for its exposure to all risk factors
- The R^2 measures the proportion of asset i's return variance that is explained by the risk factors

Example 1: Capital Asset Pricing Model (CAPM)

Theoretical Model of CAPM:

$$E[R_i] - R_f = \beta_i(E[R_m] - R_f)$$

Time-Series Regression to test the model:

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_i(R_{m,t} - R_{f,t}) + \epsilon_{i,t} \quad \text{for } t = 1, \ldots, T$$

- $R_{m,t} - R_{f,t}$: The excess return of the market portfolio at time t. This is our single risk factor ($F_{1,t}$)
- β_i: The sensitivity of asset i to movements in the market
- If the CAPM holds, the estimated intercept α_i should be zero for all assets

Example 2: Fama-French Three-Factor Model

The Fama-French model extends the CAPM by adding two additional risk factors to better explain stock returns: size and value

$$
\begin{aligned}
R_{i,t} - R_{f,t} = \alpha_i + \beta_{i,Mkt}(R_{m,t} - R_{f,t}) \\
+ \beta_{i,SMB} SMB_t \\
+ \beta_{i,HML} HML_t + \epsilon_{i,t}
\end{aligned}
$$

The Three Factors:

1. $(R_{m,t} - R_{f,t})$: The market risk premium, same as in the CAPM
2. SMB_t (Small Minus Big): The return of a portfolio of small-cap stocks minus the return of a portfolio of large-cap stocks. This captures the **size premium**.
3. HML_t (High Minus Low): The return of a portfolio of high book-to-market (value) stocks minus the return of a portfolio of low book-to-market (growth) stocks. This captures the **value premium**.

Example 1: Cross-Sectional Regression with Factor Loadings

$$\bar{R}_i^e = \gamma_0 + \gamma_1 \hat{\beta}_{i,1} + \cdots + \alpha_i \quad \text{for } i = 1, \ldots, N$$

- $\bar{R}_i^e$: The average excess return for asset i over the entire sample period
- $\hat{\beta}_{i,1}$: The estimated risk exposure (beta) of asset i, obtained from a prior time-series regression
- γ_1: The estimated **risk premium**
  - γ_1 answers: "How much extra return is earned for a one-unit increase in risk (β)?"
- α_i: The **mispricing error**
  - If an asset pricing model is correct, the α_i values should be statistically indistinguishable from zero
- The R^2 of this regression is not a primary focus; the statistical significance of the risk premium (γ_1) and the magnitude of the pricing errors (α_i) are what matter

Example 2: Cross-Sectional Regression with Firm Characteristics

Goal Explain the cross-section of expected returns using firm-level characteristics

$$\bar{R}_i^e = \gamma_0 + \gamma' Z_i + \alpha_i \quad \text{for } i = 1, \ldots, N$$

- $\bar{R}_i^e$: Average excess return of firm/asset *i* over the sample
- Z_i: Vector of characteristics, e.g. size (log market equity), book-to-market, momentum
- γ: **characteristic premia** (how expected return changes with characteristics)
- α_i: Pricing error / unexplained component for asset *i*
- In practice, characteristics are often standardized (z-scores) to ease interpretation

Panel Regression

$$R_{i,t} = \gamma' X_{i,t} + u_i + \lambda_t + \epsilon_{i,t} \quad \text{for } i = 1, \ldots, N \text{ and } t = 1, \ldots, T$$

- $R_{i,t}$: The return of asset i at time t
- $X_{i,t}$: A vector of explanatory variables, which can include both time-series factors (e.g., market return) and cross-sectional characteristics (e.g., firm size)
- γ: A vector of coefficients corresponding to the variables in $X_{i,t}$
- u_i: Firm fixed effects, which capture all unobserved, time-invariant heterogeneity specific to firm i (e.g., management quality, industry)
- λ_t: Time fixed effects, which capture any unobserved shocks in period t that affect all firms (e.g., macroeconomic news, interest rate changes)
- This approach provides a powerful way to isolate the effect of $X_{i,t}$ on returns