Intro to Time Series and Semester Review

- ***Please download:*** ***12-TheFed.xls***
- ***Presentations next week!***

3 Apr 2012
Dr. Sean Ho

busi275.seanho.com

Trinity Western University

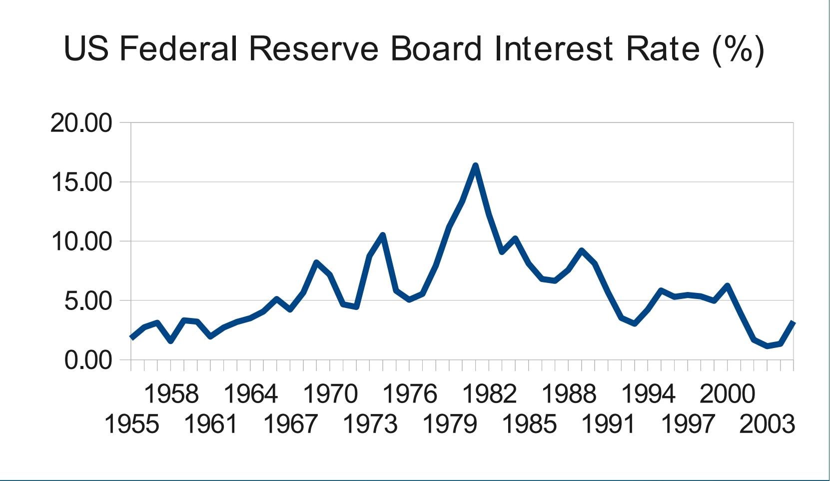
Outline for today

- Time series data: dependent observations
- Trend-based approach:
  - Trends, cycles, seasons
  - Dummy coding a seasonal model
  - Additive vs. multiplicative model
- Autoregressive approach:
  - Autocorrelation
  - Correlogram and the AR(p) model
  - Finite differencing and the ARIMA model
- Semester review

Time series data

- Time is one of the independent variables
  - Often only 1 DV and 1 IV (time)
  - But can also have other time-varying IVs
- Why not just use regression with time as the IV?
  - Assumptions of regression: in particular, observations need to be independent!
- Two (complementary) approaches:
  - Model time-varying patterns and factor them out, leaving independent (uncorrel) resids
  - Model the conditional dependence of current value on past values

Patterns / trends



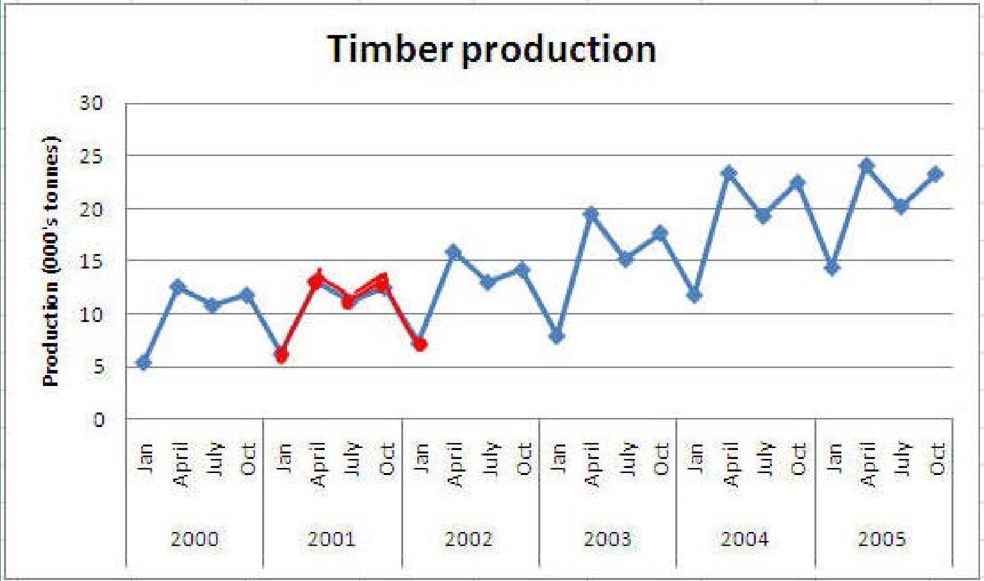
- Patterns to look for:
  - Trend: linear growth/loss
    - Or non-linear: t^λ, ln(t), S-curve, etc.
  - Cycle: multi-year repeating pattern
  - Season: pattern that repeats each year
    - e.g., if data is quarterly, use dummy vars for each season: $b_2S_2 + b_3S_3 + b_4S_4$
- Additive model:
  - $Y_t = (b_0 + b_1t)$ + (cyclical component) + (seasonal component) + (residual)
- Assumptions: residuals are independent, normally distributed, with constant variance

Seasonal pattern

t	Qtr	Sales	Sp	Su	Fa
1	2011 Wi	$18k	0	0	0
2	2011 Sp	$25k	1	0	0
3	2011 Su	$33k	0	1	0
4	2011 Fa	$23k	0	0	1
5	2012 Wi	$20k	0	0	0

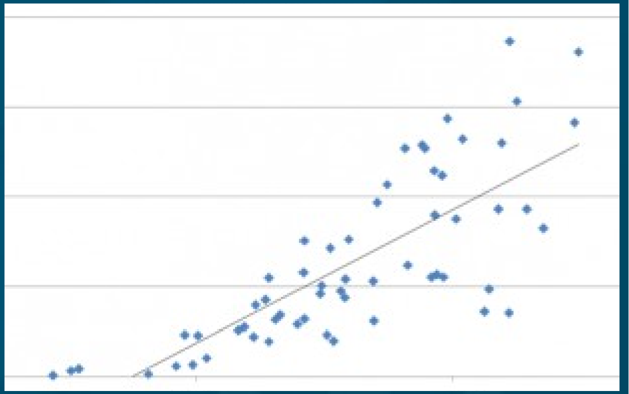
- e.g., quarterly retail sales
- Use dummy vars:
  - Pick a base case, say Wi
  - 3 dummy vars: Sp, Su, Fa
- Additive seasonal model:
  - $\hat{Y}_t = (b_0 + b_1*t) + b_2*Sp + b_3*Su + b_4*Fa$
  - (predicted) = (trend) + (seasonal)
  - 1+3 predictors
- If monthly data instead,
  - Try 11 dummy vars

Nayland College



Additive vs. multiplicative

- Homoscedasticity of residuals is often an issue
- Check residual plot: resids vs. predicted value
  - Or similar "Spread vs. level" plot: √(std resids) vs. predicted



- If you see a distinct "fan" shape,
  - i.e., the SD of residuals grows with the level of the variable,
- Then apply a log transform to the variable:
  - $\ln(Y_t)$ = (linear) + (cyclic) + (seasonal)
- This is equivalent to a multiplicative model:
  - Y_t = (linear) * (cyclic) * (seasonal)

Outline for today

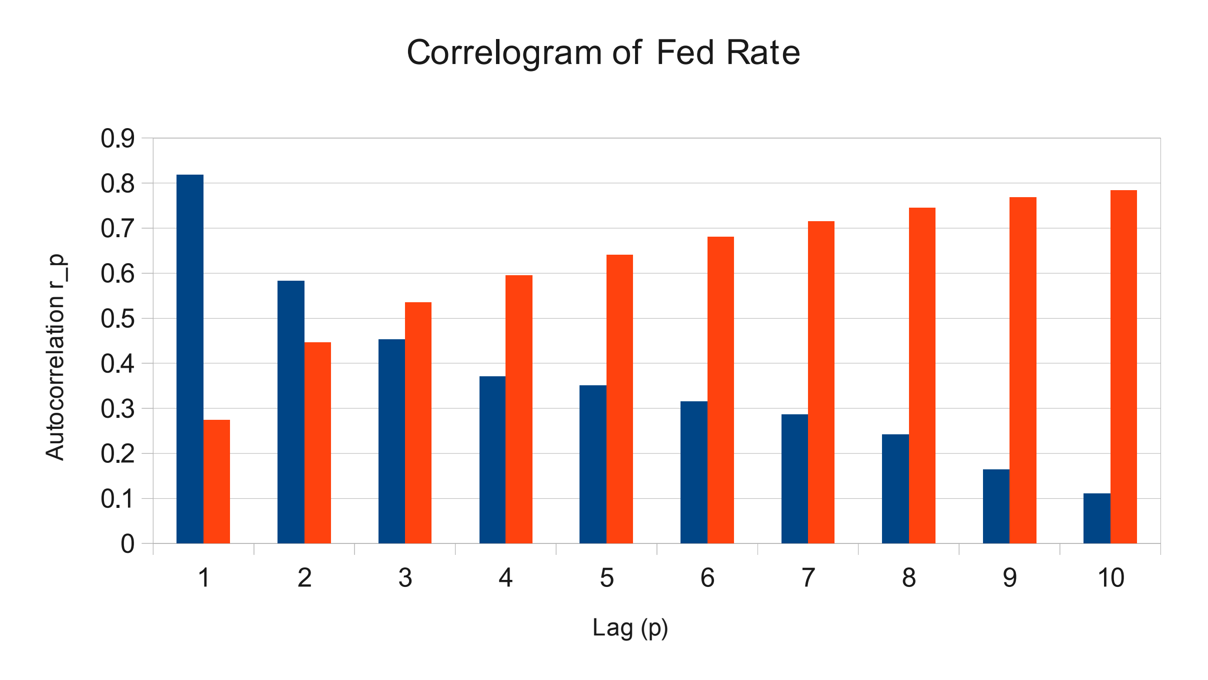
- Time series data: dependent observations
- Trend-based approach:
  - Trends, cycles, seasons
  - Dummy coding a seasonal model
  - Additive vs. multiplicative model
- Autoregressive approach:
  - Autocorrelation
  - Correlogram and the AR(p) model
  - Finite differencing and the ARIMA model
- Semester review

Autocorrelation

- Another approach models the correlation of the current value against past values:
  - P(Y_t | Y_{t-1})
  - Or in general: P(Y_t | {Y_s: all $s<t$})
- The autocorrelation (ACF) r_p of a variable Y is the correlation of the variable against a time-shifted version of itself:
  - p is the lag (always positive): number of time units to shift
  - p.629 #14-64c: (sales) vs. (ad in prev wk)
- e.g., quarterly seasonal data may have large r_4

Correlogram and AR model

- The correlogram is a column chart illustrating the autocorrelation for various lags
- Statistical software will also show the critical value for each autocorrelation
  - Autocorrelations that are significant suggest an autoregressive model with lag p: AR(p)
- TheFed data: AR(2) model
- Current rate depends on prev 2 years (“memory”)



ARIMA model

- ARIMA model combines three parts:
  - AR (autoregression) uses recent values to predict current value
  - MA (“moving average”) uses recent errors in prediction (residuals) to predict current resid
  - I (“integration”) uses finite differencing to factor out consistent trends
    - Looks at $Y_t - Y_{t-d}$, where d is the lag
    - Year-over-year change (annual data): d=1
    - Year-over-year change (quarterly): d=4
- The Box-Jenkins method is a way to find the 3 lags that parameterise an ARIMA(p,d,q) model

Combining approaches

- The trend-based approach and the autoregressive approach can be combined:
- First fit broad trends/cycles/seasons
  - Resulting residuals (de-trended, de-seasonalized data) may still be auto-correlated
- Use correlograms to choose an ARIMA model for the residuals
- Goal is to get the residuals to be small, independent, normally distributed, and with constant variance

Outline for today

- Time series data: dependent observations
- Trend-based approach:
  - Trends, cycles, seasons
  - Dummy coding a seasonal model
  - Additive vs. multiplicative model
- Autoregressive approach:
  - Autocorrelation
  - Correlogram and the AR(p) model
  - Finite differencing and the ARIMA model
- Semester review

Overview: foundation

- Intro: variables, sampling (Ch1)
- Exploring data:
  - Via charts (Ch2), via descriptives (Ch3)
- Probability and independence (Ch4)
- Probability distributions:
  - Discrete: binom, Poisson, hypg (Ch5)
  - Continuous: norm, unif, expon (Ch6)
- Sampling distributions (Ch7, 8)
  - SDSM (norm and t-dist), binomial
  - Types of problems: % area, conf. int., n
- Hypothesis testing (Ch9):
  - H_0/H_A, rej / fail rej, Type-I/II, α/β, p-value

Overview: statistical tests

- *T*-tests (Ch10):
  - 1 sample mean (ch9)
  - Two independent samples (het σ, hom σ)
  - Paired data (Excel type 1)
- Regression (Ch14-15):
  - Linear model, predicted ŷ, residuals
  - R^2, F-test, t-test on slopes, interaction
- ANOVA (ch12):
  - One-way + Tukey-Kramer
  - Blocking (w/o repl) + Fisher's LSD
  - Two-way (w/repl), interaction
- χ^2 (ch13): contingency tables, O vs. E

Ch7-8: Sampling distributions

- Sampling distributions:
  - SDSM, w/σ: NORMDIST(), SE = $\sigma/\sqrt{n}$
  - SDSM, w/s: TDIST(), SE = $s/\sqrt{n}$
  - Binomial proportion: norm, SE = $\sqrt{(pq / n)}$
- Types of problems: area, μ, thresh, n, σ
  - Area: prob of getting a sample in given range
  - Threshold: e.g., confidence interval
  - n: minimum sample size

Ch9: Hypothesis testing

- Decision making
- H_0 vs. H_A , in words and notation (e.g., $\mu_1 \neq \mu_2$)
- Conclusions: reject H_0 vs. fail to reject H_0
- Risks/errors: Type-I vs. Type-II
  - Level of significance: α
  - Power: $1-\beta$
- p-value: what is it, how do we use it?

Ch10: t-tests

- T-test on 1 sample (ch8-9):
  - SDSM: $SE = s/\sqrt{n}$
  - Binomial proportions: $SE = \sqrt{(pq/n)}$
- T-test on two independent samples, general:
  - $SE = \sqrt{(SE_1^2 + SE_2^2)}$, df = complicated
- T-test on two independent samples, similar σ:
  - $SE = s_p \sqrt{(1/n_1 + 1/n_2)}$, $df = df_1 + df_2$
- T-test on two proportions:
  - $SE = \sqrt{(SE_1^2 + SE_2^2)}$, use z instead of t
- T-test on paired data:
  - $SE = s_d / \sqrt{n}$, df = (#pairs) – 1

Ch14: Regression

- Scatter plots and correlation, t-test on r
  - R^2 and % variability explained
- Linear model $Y = b_0 + b_1X + \varepsilon$
  - Finding+interpreting slope+intercept
  - Finding+interpreting s_ε (STEYX)
- Assumptions / diagnostics:
  - Linearity + homoscedasticity (residual plots)
  - Normality of residuals (histogram)
  - *(skip: non-collinearity + indep of resids)*
- ch15: only concepts of multiple regression, especially moderation

Ch12-13: Categorical data

- Ch12: ANOVA:
  - H_0 / H_A, global F-test, concept of follow-up
  - One-way ANOVA + Tukey-Kramer
  - Blocking ANOVA + Fisher's LSD
    - F-test for main factor effect
    - F-test for whether blocking is needed
  - Two-way ANOVA
    - F-test for each main effect
    - F-test for interaction
- Ch13: χ^2 (O vs. E)
  - 1 var vs. uniform, normal
  - 2 vars (contingency table): independence

TODO

- Presentations next week
  - Remember your potential clients:
    what questions would they like answered?
  - Tell a story/narrative in your presentation
- Email me your preferences (if any) for time slot
  - I will post the schedule tomorrow
- You will be writing feedback to each group
  - Short answer form, on myCourses
- Upload or share your presentation slides
  - Can be done shortly after your presentation
- Paper is due 16Apr, final exam is 26Apr 9am