Using Time Series Analysis to Estimate and Forecast the Effect of Tobacco Control Funding on Smoking Behavior and Health Care Expenditure

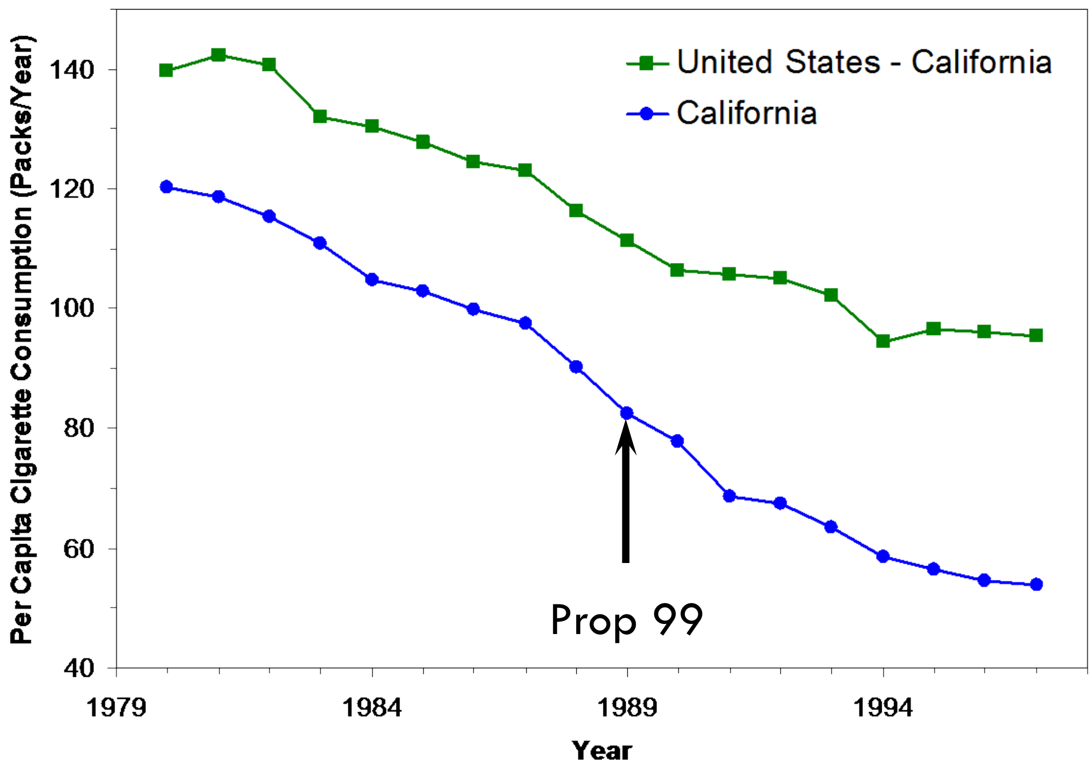
Jim Lightwood, UCSF

Origin of research question

- Research Question
  - Nationally representative individual survey and model-based estimates of the effects of smoking show that it should have a major impact on health care expenditures
    - So, if we compare populations with different histories of smoking behavior, shouldn't we see differences in
      - Adverse events
      - Health care resource utilization
      - Per capita health care expenditures

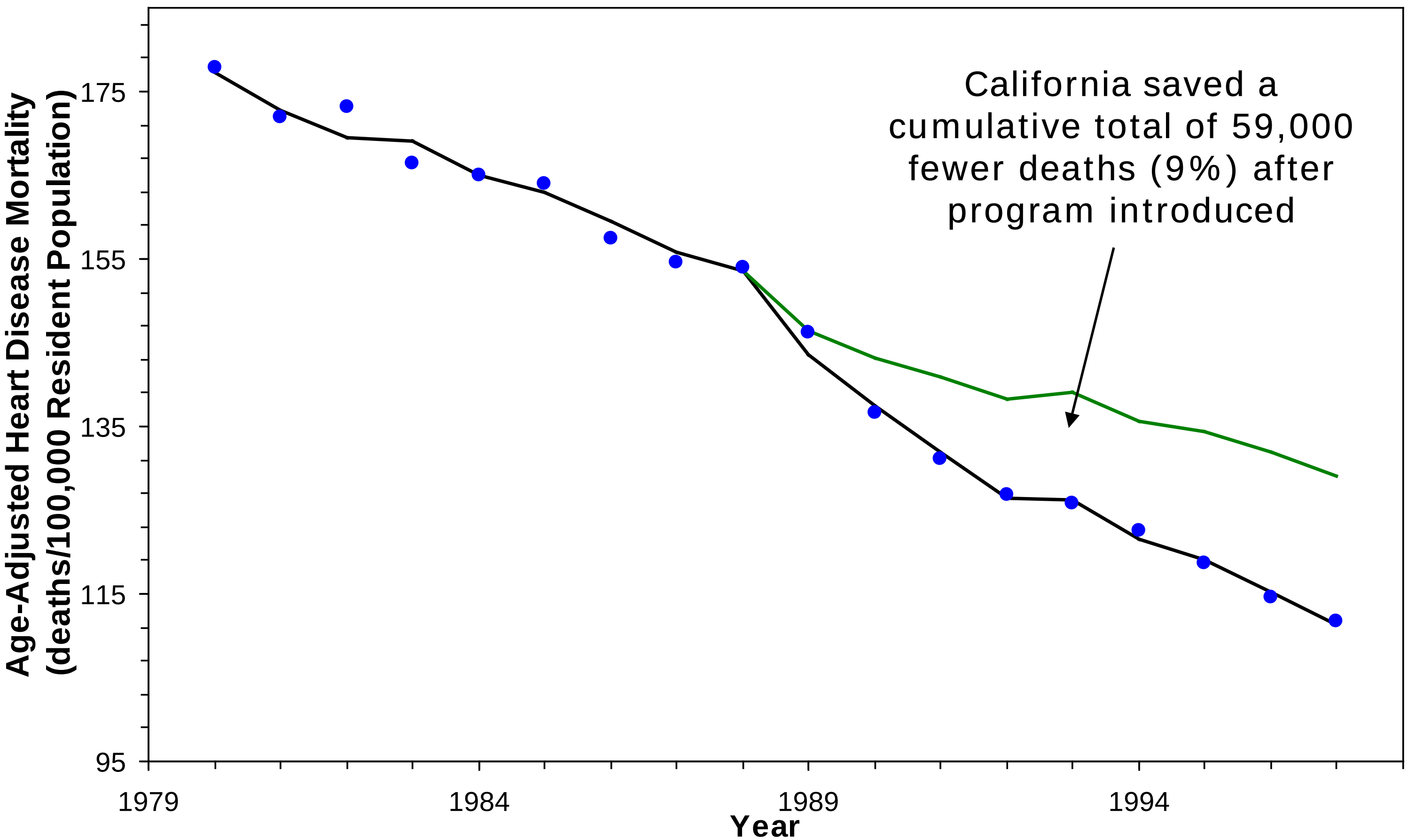
    in different populations due to changes in smoking behavior in aggregate data?

Cigarette consumption dropped in California compared to rest of US



Source: Fichtenberg and Glantz (2000)

Heart disease deaths dropped in California faster after, compare to before, Prop 99



Source: Fichtenberg and Glantz (2000)

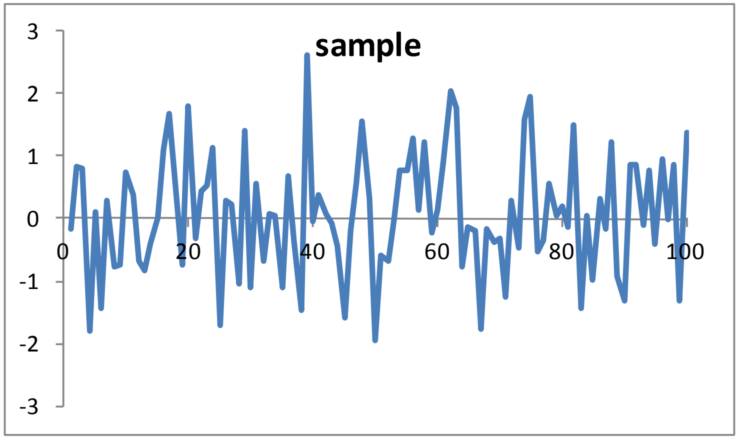
Two common types of time series

- Covariance Stationarity
  - Mean and variance do not vary over time

- Unit root nonstationarity
  - Variance goes to infinity as time increases (in irregular jumps for 'small' samples)
  - Mean is undefined
    - In terms of sample properties, the sample of mean of an initial segment of observations tells you nothing about the next observation
- Statistical analysis requires that a constant mean be estimated and variance be finite
  - So doing statistics on a nonstationary time series can lead to spurious results

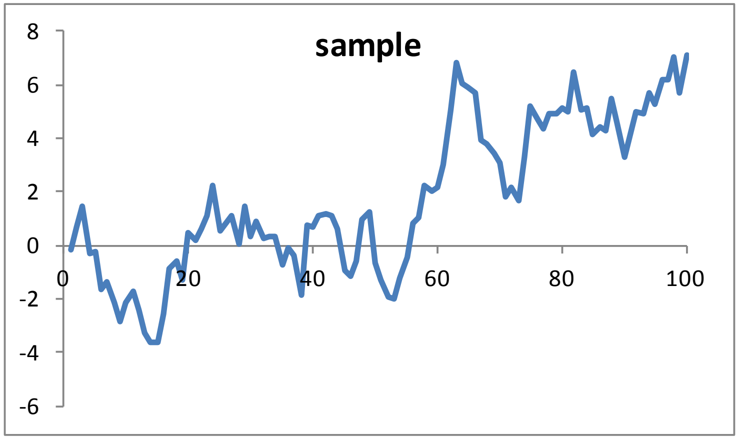
Sample Path of stationary versus unit root nonstationary process for 100 observations

Stationary



Innovations at each time are unit normal

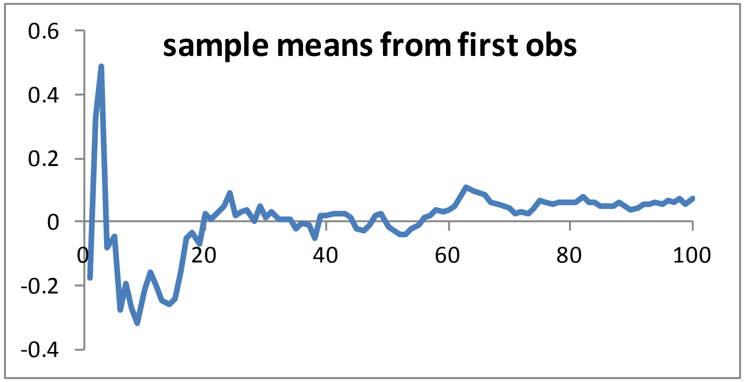
Unit root nonstationary



Note: in small to moderate samples, similar pattern exists with stationary series with high persistence

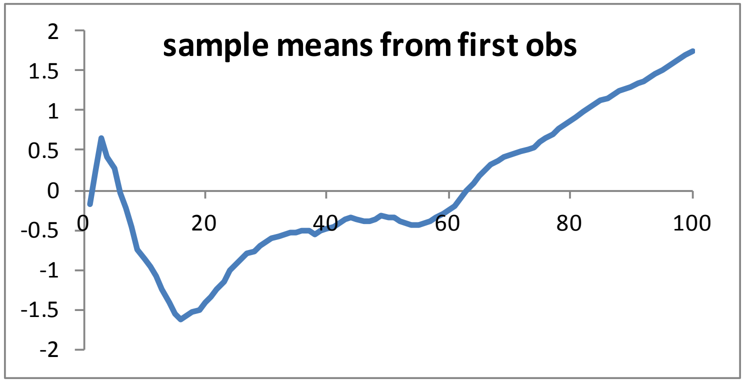
Mean of sample path of stationary versus unit root nonstationary process for 100 observations

Stationary



Innovations at each time are unit normal

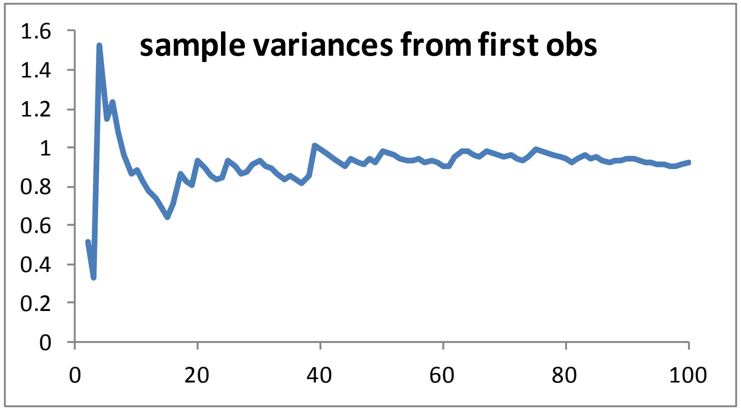
Unit root nonstationary



Note: in small to moderate samples, similar pattern exists with stationary series with high persistence

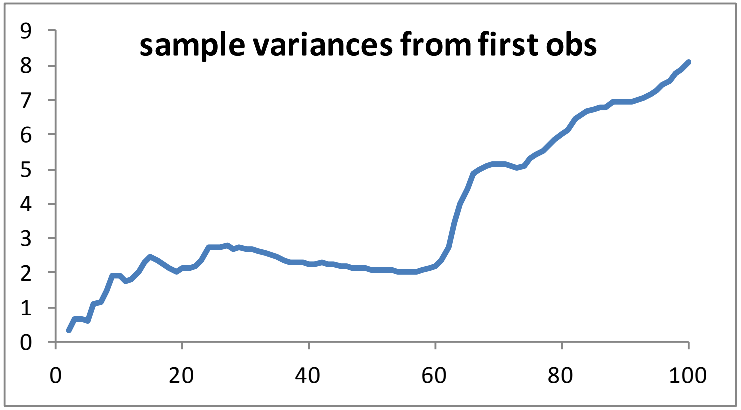
Variance of sample path of stationary versus unit root nonstationary process for 100 observations

Stationary



Innovations at each time are unit normal

Unit root nonstationary

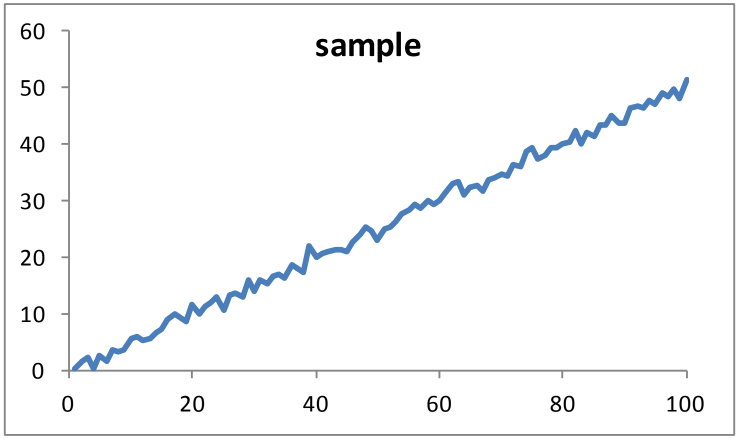


Note: in small to moderate samples, similar pattern exists with stationary series with high persistence

Unit root process can be generalized to trending series. Sample Path of stationary versus unit root nonstationary trending series

Stationary trend

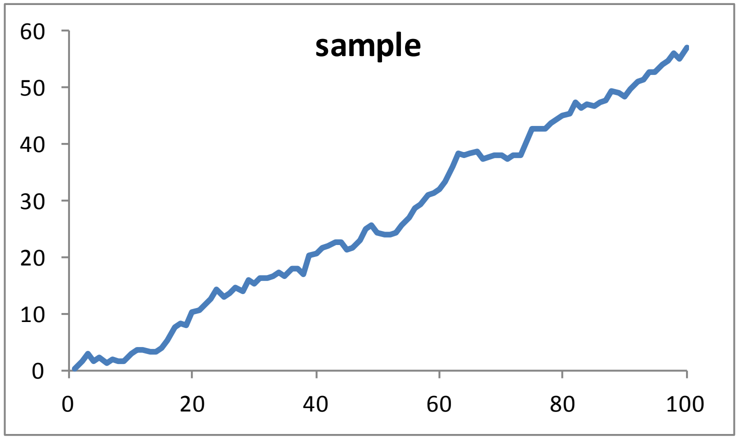
$$y_t = 0.5 * time + \varepsilon_t$$



Innovations at each time are unit normal

Unit root nonstationary series with drift

$$y_t = 0.5 + y_{t-1} + \varepsilon_t$$



Note: in small to moderate samples, similar pattern exists with stationary series with high persistence

How to analyze relationships between two unit root nonstationary variables

- Suppose two time series, *x* and *y* are unit root nonstationary
- For a long time, it was thought impossible to analyze relationship between two such variables
- Because the error term would inherit the unit root nonstationarity
- Cannot use standard statistical procedures if the error term does not have a well defined mean and has infinite variance
  - Which seems to be the case for total and per capita health care expenditures (which act more like macroeconomic and financial time series than epidemiological time series)

Spurious time series regression

- Suppose you have a regression between *y* and *x*

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

- And the error term, ε, is unit root non-stationary
- Problems ensue with statistical inference on the regression coefficient β
  - β will not converge to a constant, but to a random walk
  - β will be different from zero, using standard t-test, more than half the time in small to moderate samples, even if *y* and *x* are independent time series
  - Problems cannot be fixed by adjusting regression to make error term stationary (e.g., Cochrane-Orcutt estimator)

Long and short run relationships

- The cointegrating regression

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

gives the long run relationship between y and x, after all short run adjustment processes die out

- Short run adjustment process can be estimated in a separate step using residuals from the cointegrating regression (the 'error' or 'equilibrium' correction model)

$$\Delta y_t = \gamma_0 + \gamma_1 \varepsilon_{t-1} + \gamma_2 \Delta y_{t-1} \gamma_3 + \Delta x_{t-1} + v_t$$

Nice things about cointegrating regressions

- The regression coefficients will be consistent under wide range of error processes
  - Stationary error term is only requirement
- Most common situations, the slope coefficient converges to asymptotic value at rate of number of observations, rather than square root
- In moderate size to large samples (and under certain assumptions, in small samples), endogeneity bias from RHS variable is not a problem
  - But, new instrumental variables techniques eliminate problem in all but smallest sample sizes

Nice things about cointegrating regressions (cont.)

- Fact that long run relationship must be a static linear regression helps narrow specification search
- Unlike traditional approaches to nonstationary time series (e.g. ARIMA), does not 'throw away' information in long run relationships
- Can use on aggregate data, often cheaper and easier to acquire, for forecasting and prediction
- In small to moderate size samples, works well when data are stationary but have high autocorrelation

Not so nice things about cointegrating regressions

- Must be careful to verify that residuals are stationary
  - Particularly, serial correlation in residuals is not too high
  - May be difficult in small to moderately sized samples
- Theory and methods for nonlinear long run relationships are only now under development
- When more than two variables, and there is more than one cointegrating relationship ALL regression coefficients may be inconsistent (that is, results in a severe identification problem)

Difficulty of estimating effects of tobacco control program on health care expenditure

- Many factors (some unobservable) can change
  - smoking behavior
    - Social attitudes toward smoking
    - Behavior of tobacco industry
  - Per capita health care expenditures
    - Other health risks (obesity, drinking, blood pressure…)
    - Technological progress, changes in health finance policy, insurance and health care industry structure
- The changes in these factors persist over time

Solution to unobserved common trends

- Time series evidence indicates that the observable variables in California and other states
  - Follow common trends across regions and states
  - Patterns over time give information about how to write down the mathematical model for the long run relationships
- Average values of explanatory variables for control populations that model unobservable trends can be included in the model
  - Choice of control population does not make noticeable difference in estimates of long run relationships

First California Tobacco Control → Smoking → Health Care Expenditure Model

- Effect of cumulative per capita tobacco control funding on per capita cigarette consumption

$$cigarette\_consumption_{CA,t} = \alpha_0 + \alpha_1 cigarette\_consumption_{Control,t} + \alpha_2 (TC\_expenditure_{CA,t} - TC\_expenditure_{Control,t}) + \alpha_3 price_{CA,t} + \alpha_4 price_{Control,t} + \alpha_4 (t - 1980) + \varepsilon_t$$

- Effect of per capita cigarette consumption on per capita health care expenditures:

$$healthcare\_expenditure_{CA,t} = \gamma_0 + \gamma_1 healthcare\_expenditure_{Control,t} + \gamma_2 (cigarette\ consumption_{Control,t} - cigarette\ consumption_{CA,t}) + \upsilon_t$$

Include a measure of common trends across states. Can be specially chosen set of control states, or simple cross sectional average of all states

Sample period: 1980 to 2004: 25 observations
Health care expenditure: per capita per year
cigarette consumption: packs per capita per year
TC_expenditure: *cumulative* per capita tobacco control expenditure

Below are Published Results in PLoS Medicine

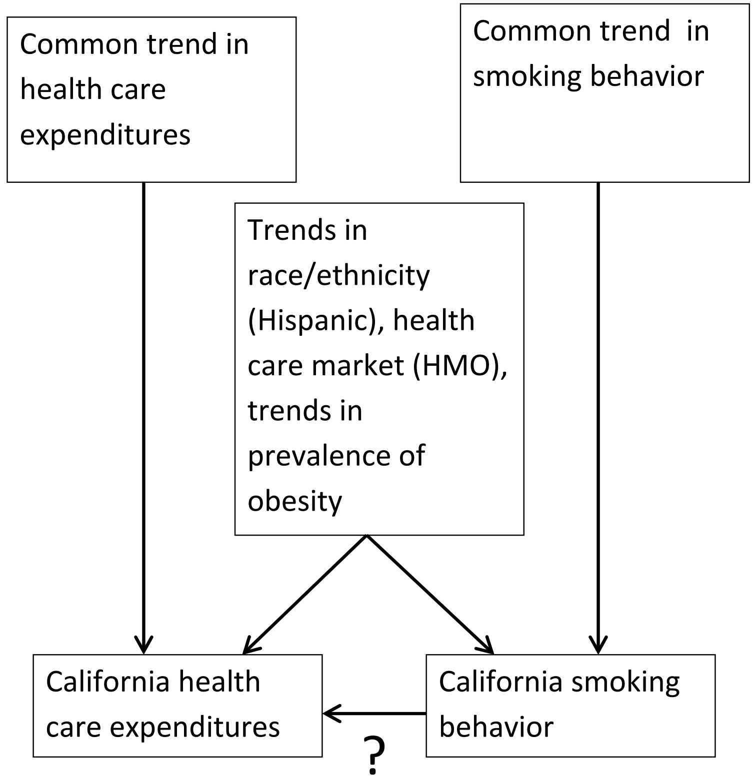
Dependent Variable	Equation	Results	n	R^2	RMSE	Autocorrelation
California per capita personal health care expenditures (2004$)	Cointegrating regression	$h_{CA,t} = \$2{,}736\ (\pm\ \$173) + 0.599\ (\pm\ 0.0519)\ h_{control,t} - \$27.00\ (\pm\ \$1.82)(/\textit{pack per capita})\ (s_{control,t} - s_{CA,t}) + \upsilon_{1,t}$	25	0.91	46.0	0.09
	Equilibrium correction model	$\Delta h_{CA,t} = -0.759\ (\pm\ 0.390)\ \upsilon_{1,t-1} + 0.481\ (\pm\ 0.221)\ \Delta h_{CA,t-1} + \varepsilon_{1,t}$	23	0.21	71.9	0.11
Difference in cigarette consumption in California and control states (packs per capita)	Cointegrating regression	$(s_{control,t} - s_{CA,t}) = 30.3\ (\pm\ 2.15) + 0.261\ (\pm\ 0.0780)\ (\textit{packs per capita})/(\textit{\$ per capita})\ (E_{CA,t} - E_{control,t}) + 11.3\ (\pm\ 2.20)\ (\textit{packs per capita})/(\textit{\$ per pack})\ p_{CA,t} - 22.6\ (\pm\ 2.90)\ (\textit{packs per capita})/(\textit{\$ per pack}) p_{control,t} + 1.69\ (\pm\ 0.187)\ (\textit{packs per capita/year})\ (t - 1980) + \upsilon_{2,t}$	25	0.98	1.75	−0.23[a]
	Equilibrium correction model	$\Delta(s_{control,t} - s_{CA,t}) = 0.946\ (\pm\ 0.404) - 0.960\ (\pm\ 0.232)\ \upsilon_{2,t-1} + 0.315\ (\pm\ 0.185)\ \Delta(s_{control,t-1} - s_{CA,t-1}) + \varepsilon_{2,t}$	23	0.46	1.57	−0.04

Source: Lightwood, Dinno and Glantz 2008

Important to check that residuals are stationary: do not have high autocorrelation

Cautions in use of cointegration and common trends in modeling

- Must be watchful for possibility of an unknown trend driving all variables in the California
  - Income, Hispanic population, prevalence of obesity
- So, need extensive sensitivity analysis to search for other trends that may produce spurious causal interpretation attributed to an estimated association, and rule them out
- Should have pre-existing causal model established, or suggested, by clinical research, randomized controlled trials, or analysis of individual survey data (e.g. MEPS, NHIS, etc.)



References

Examples by presenter

Lightwood, J. and S. Glantz (2011a). Predicted Effect of California Tobacco Control Funding on Smoking Prevalence, Cigarette Consumption and Healthcare Costs, 2012-2016. Oakland, CA, California Tobacco Related Disease Research Program.

Lightwood, J. and S. Glantz (2011b). Predicted Effect of California Tobacco Control Funding on Smoking Prevalence, Cigarette Consumption, and Healthcare Costs, 2012-2016. San Francisco, CA, Center for Tobacco Control Research and Education, University of California, San Francisco.

Lightwood, J. and S. Glantz (forthcoming 2012). "The Effect of the California Tobacco Control Program on Smoking Prevalence, Cigarette Consumption, and Healthcare Costs: 1985-2008." PLoS ONE forthcoming.

Lightwood, J. M., A. Dinno, et al. (2008). "Effect of the California tobacco control program on personal health care expenditures." PLoS Medicine **5**(8): e178

Introduction to time series, including stationary approaches and cointegration

Enders, W. (2004). Applied Econometric Time Series. Hoboken, NJ, John Wiley and Sons.

Maddala, G. S. and I.-M. Kim (1998). Unit Roots, Cointegration, and Structural Change. Cambridge, Cambridge University Press.

Use of cointegrating techniques with stationary but highly persistent data

Hoover, K., S. Johansen, et al. (2008). "Allowing the data to speak freely: the macroeconometrics of the cointegrated autoregression." American Economic Review: Papers and Proceedings **98**: 251-255.

Introduction to modeling with common trends

Breitung, J. and M. Pesaran (2008). Unit roots and cointegration in panels. The Econometrics of Panel Data. L. Matyas and P. Sevestre. Berlin, Springer-Verlag: 279-322.

Ebarhardt, M. (2009). Nonstationary Panel Econometrics and Common Factor Models: An Introductory Reader. Oxford, UK, Department of Economics, Oxford University: 1-60.