Lecture 1
Intro to Spatial and Temporal Data

Dennis Sun
Stanford University
Stats 253

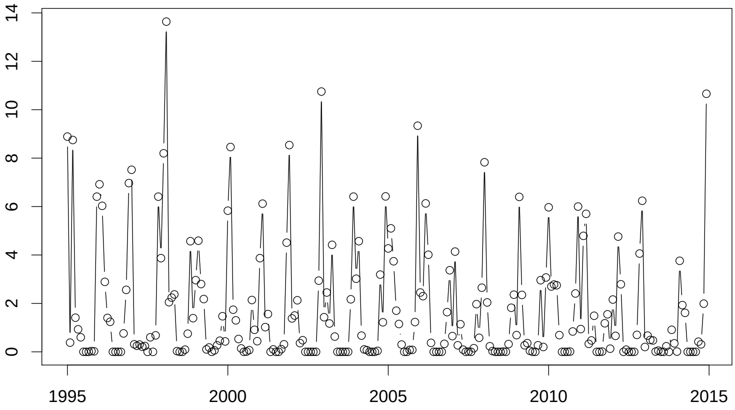
June 22, 2015

1. What is Spatial and Temporal Data?
2. Trend Modeling
3. Omitted Variables
4. Overview of this Class

1. What is Spatial and Temporal Data?
2. Trend Modeling
3. Omitted Variables
4. Overview of this Class

Temporal Data

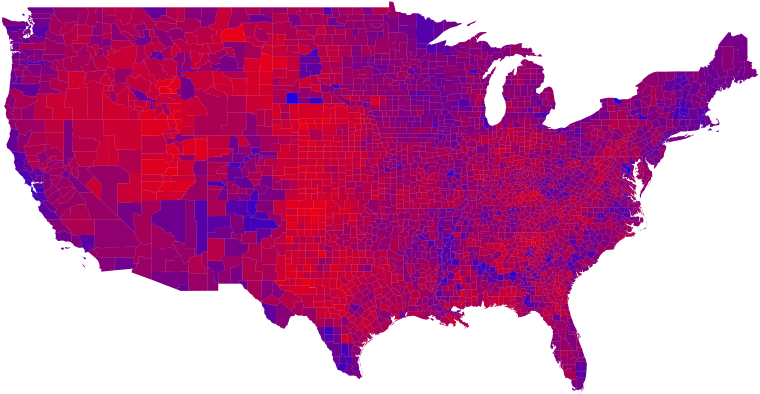
Temporal data are also called **time series**.



Monthly Rainfall in San Francisco

Spatial Data

Spatial observations can be **areal units**...



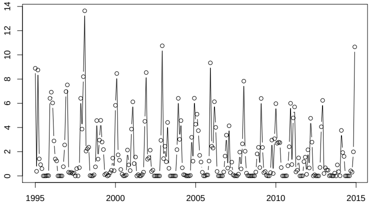
Percent of votes for George W. Bush in 2004 election.

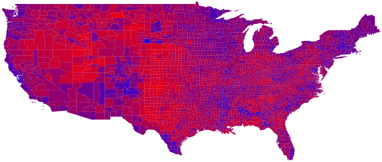
Spatial Data

...or **points in space**.

San Jose house prices from zillow.com

What do the two have in common?





Observations that are close in time or space are similar.

Why is this the case?

Common or similar factors drive observations that are nearby in time and space.

- The meteorological phenomena that drive rainfall (e.g., El Niño) in one month typically lasts a few months.
- Religion and race are strong predictors of voters' choices. These are likely to be similar in nearby regions.
- School quality is a strong predictor of house prices. Nearby houses belong to the same school district.

To make this precise, assume that each observation y_i can be modeled as a function of predictors $\mathbf{x}_i$:

$$y_i = \underbrace{f(\mathbf{x}_i)}_{\text{trend}} + \underbrace{\epsilon_i}_{\text{noise}}$$

Linear Models

- We will focus on the most common model for the trend, a **linear model**:

$$f(\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta},$$

  although there are others (loess, splines, etc.).

- We estimate $\boldsymbol{\beta}$ by **ordinary least squares** (OLS)

$$\begin{aligned}\hat{\boldsymbol{\beta}} &\stackrel{def}{=} \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \\ &= \operatorname*{argmin}_{\boldsymbol{\beta}} ||\mathbf{y} - X\boldsymbol{\beta}||^2 \\ &= (X^T X)^{-1} X^T \mathbf{y}\end{aligned}$$

- Is this a good estimator?

Properties of OLS

If we assume that $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathrm{E}[\boldsymbol{\epsilon}|X] = \mathbf{0}$, then

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (X^T X)^{-1} X^T \mathbf{y} \\ &= (X^T X)^{-1} X^T (X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta} + (X^T X)^{-1} X^T \boldsymbol{\epsilon}.\end{aligned}$$

Then, $\mathrm{E}[\hat{\boldsymbol{\beta}}|X] = \boldsymbol{\beta} + \mathrm{E}[(X^T X)^{-1} X^T \boldsymbol{\epsilon}|X] = \boldsymbol{\beta}$, so the OLS estimator is unbiased.

In fact, it is the "best" linear unbiased estimator. (More on this next time.)

Example: House Prices in Florida

```
Call:
lm(formula = price ~ size + beds + baths + new, data = houses)

Residuals:
     Min       1Q   Median       3Q      Max
-215.747  -30.833   -5.574   18.800  164.471

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.84922   27.26116  -1.058  0.29262
size          0.11812    0.01232   9.585 1.27e-15 ***
beds         -8.20238   10.44984  -0.785  0.43445
baths         5.27378   13.08017   0.403  0.68772
new          54.56238   19.21489   2.840  0.00553 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54.25 on 95 degrees of freedom
Multiple R-squared:  0.7245, Adjusted R-squared:  0.713
F-statistic: 62.47 on 4 and 95 DF,  p-value: < 2.2e-16
```

Where do the standard errors come from?

If we further assume $\mathrm{Var}[\boldsymbol{\epsilon}|X] = \sigma^2 I$, then we can calculate:

$$\begin{aligned}
\mathrm{Var}[\hat{\boldsymbol{\beta}}|X] &= \mathrm{Var}\left[\boldsymbol{\beta} + (X^T X)^{-1} X^T \boldsymbol{\epsilon}|X\right] \\
&= \left((X^T X)^{-1} X^T\right) \mathrm{Var}\left[\boldsymbol{\epsilon}|X\right] \underbrace{\left((X^T X)^{-1} X^T\right)^T}_{X(X^T X)^{-1}} \\
&= \sigma^2 \left((X^T X)^{-1} X^T\right) \left(X(X^T X)^{-1}\right) \\
&= \sigma^2 (X^T X)^{-1}.
\end{aligned}$$

Since $\hat{\boldsymbol{\beta}}$ is a random vector, this is a **covariance matrix**:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix}
\mathrm{Var}(\hat{\beta}_1) & \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \dots & \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_p) \\
\mathrm{Cov}(\hat{\beta}_2, \hat{\beta}_1) & \mathrm{Var}(\hat{\beta}_2) & \dots & \mathrm{Cov}(\hat{\beta}_2, \hat{\beta}_p) \\
\vdots & \vdots & \ddots & \vdots \\
\mathrm{Cov}(\hat{\beta}_p, \hat{\beta}_1) & \mathrm{Cov}(\hat{\beta}_p, \hat{\beta}_2) & \dots & \mathrm{Var}(\hat{\beta}_p)
\end{pmatrix}.$$

The square root of the diagonal elements give us the standard errors, i.e., $SE(\hat{\beta}_j) = \sqrt{\mathrm{Var}(\hat{\beta}_j)}$.

What happens if we omit a variable?

- Suppose the following model for house prices is correct:

$$\text{price}_i = \underbrace{\beta_0 + \beta_1 \cdot \text{size}_i + \beta_2 \cdot \text{new}_i}_{\text{trend}} + \underbrace{\epsilon_i}_{\text{noise}},$$

  where $\mathrm{E}[\boldsymbol{\epsilon}|\text{size}, \text{new}] = \mathbf{0}$ and $\mathrm{Var}[\boldsymbol{\epsilon}|\text{size}, \text{new}] \propto I$.

- Suppose we don't actually have data about whether a house is new or not.
- We omit it from our model, so new becomes part of the noise.

$$\text{price}_i = \underbrace{\beta_0 + \beta_1 \cdot \text{size}_i}_{\text{trend}} + \underbrace{\beta_2 \cdot \text{new}_i + \epsilon_i}_{\text{noise}},$$

  Is this a problem?

- We are fine as long as

$$\mathrm{E}[\text{noise} \mid \text{size}] = \mathbf{0} \qquad \mathrm{Var}[\text{noise} \mid \text{size}] \propto I$$

Omitted Variable Bias

Suppose the first condition is violated, i.e., $\mathrm{E}[\text{noise} \mid \text{size}] \neq 0$, i.e.,

$$\mathrm{E}[\beta_2 \cdot \text{new} + \boldsymbol{\epsilon} \mid \text{size}] \neq \mathbf{0}.$$

Since $\mathrm{E}[\boldsymbol{\epsilon} \mid \text{size}] = \mathbf{0}$, this means

$$\mathrm{E}[\beta_2 \cdot \text{new} \mid \text{size}] \neq \mathbf{0}.$$

Two things have to happen for this situation to occur:

- $\beta_2 \neq 0$: The omitted variable is relevant for predicting the response.
- $\mathrm{E}[\text{new} \mid \text{size}] \neq 0$: The omitted variable is correlated with a predictor in the model.

Omitted variables are also called **confounders**.
Since $\mathrm{E}[\text{noise} \mid \text{size}] \neq 0$, $\hat{\beta}_1$ is no longer unbiased for β_1.

Correlated Noise

- Suppose we are reasonably convinced that new is not correlated with size in our dataset.
- So we will be able to obtain an unbiased estimator for the effect of size on house prices.
- But in order for the standard errors to be valid, we need

$$\mathrm{Var}[\beta_2 \cdot \mathrm{new} + \boldsymbol{\epsilon} \,|\, \mathrm{size}] \propto I.$$

- This depends on whether

$$\mathrm{Var}[\mathrm{new} \,|\, \mathrm{size}] \propto I,$$

but chances are:

$$\mathrm{Cov}[\mathrm{new}_i, \mathrm{new}_j \,|\, \mathrm{size}] \neq 0.$$

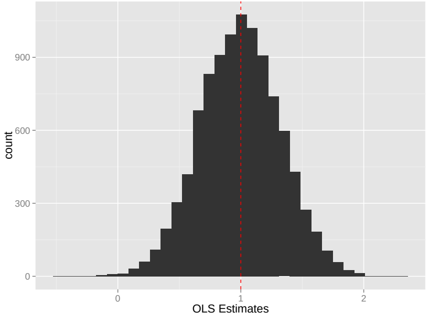
A Simulation Study

Suppose we have $n = 20$ observations from

$$y_t = \beta x_t + \epsilon_t, \ \ \beta = 1$$

where ϵ_t is correlated (generated from an AR(1) process).

Here are the OLS estimates $\hat{\beta}$ obtained over 10000 simulations.



According to the simulations:

$\mathrm{E}[\hat{\beta}|\mathbf{x}] \approx 1$, so $\hat{\beta}$ is unbiased.

$SE[\hat{\beta}|\mathbf{x}] \approx .15.$

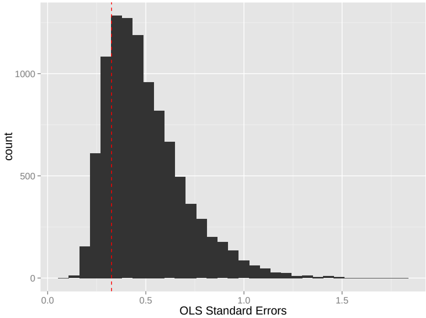
A Simulation Study

Suppose we have $n = 20$ observations from

$$y_t = \beta x_t + \epsilon_t, \;\; \beta = 1$$

where ϵ_t is correlated (generated from an AR(1) process).

Here are the naive SEs from calling the `lm` function in `R`.



OLS does not estimate the standard error appropriately.

Why study spatial and temporal statistics?

- The focus of this class will be **supervised learning**

$$y_i = f(\mathbf{x}_i) + \epsilon_i$$

  when the error is correlated.
- We will assume that the omitted variables do not lead to bias ($\mathrm{E}[\boldsymbol{\epsilon}|X] = 0$).
- If the omitted variables all have a spatial or temporal structure, then we can try to model it explicitly:

$$\mathrm{Cov}[\epsilon_i, \epsilon_j|X] = g(d(i,j)).$$

- This will allow us to (1) obtain correct inferences for the variables in the model and (2) obtain a more efficient estimator than the OLS estimator.

Course Requirements

- We'll have 3 homeworks, which will be coding / data analysis.
- We'll also have 3 in-class quizzes, which will go over the conceptual issues.
- These will be graded on a check / resubmit basis.
- For those taking the class for a letter grade, the grade will be based primarily on a final project.

Structure of the Class

- This class will meet Monday, Wednesday, Friday at 2:15pm for the first four weeks.
- The last four weeks will be dedicated to your final project. I will schedule individual meetings with students, and there may be sporadic lectures covering topics of interest to the class.

Course Website

- The course website is `stats253.stanford.edu`.
- All materials (syllabus, lecture slides, homeworks) will be posted here.
- All homework will be submitted through this course website.