ANALYSIS OF PANEL DATA

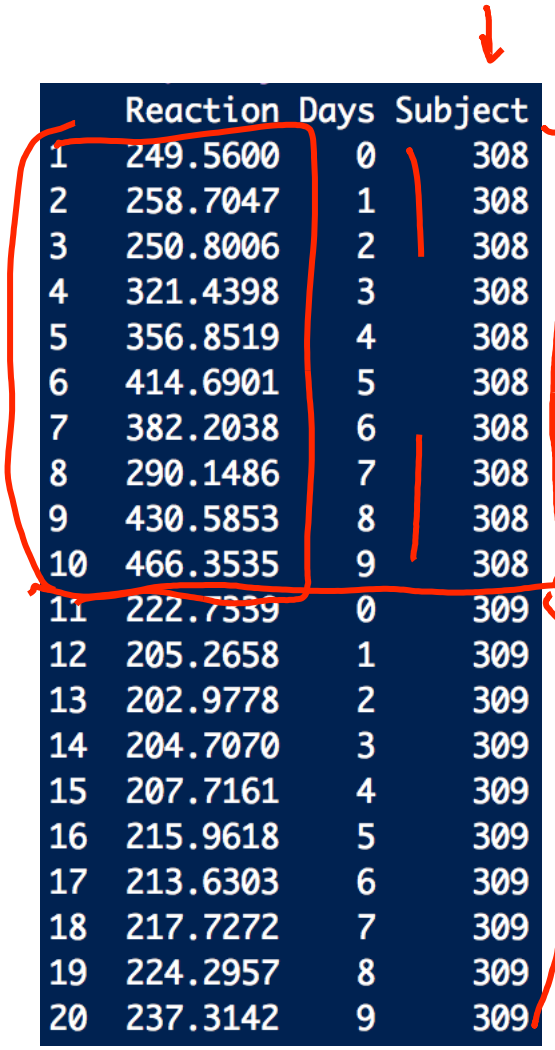
An Introduction

datascience@berkeley

Introduction to Panel Data

What Is Panel Data?

- Panel data, which are often referred to as longitudinal data, have both cross-section and time series dimensions.
- Panel data can be created by sampling the same individuals, families, departments within a company, companies, schools, cities, counties, and so on, over time.
- This gives us both the characteristics and response of interest over multiple time points.
- The example shows two individuals observed daily over a 10-day period.



	Reaction	Days	Subject
1	249.5600	0	308
2	258.7047	1	308
3	250.8006	2	308
4	321.4398	3	308
5	356.8519	4	308
6	414.6901	5	308
7	382.2038	6	308
8	290.1486	7	308
9	430.5853	8	308
10	466.3535	9	308
11	222.7339	0	309
12	205.2658	1	309
13	202.9778	2	309
14	204.7070	3	309
15	207.7161	4	309
16	215.9618	5	309
17	213.6303	6	309
18	217.7272	7	309
19	224.2957	8	309
20	237.3142	9	309

Potentials and Capabilities

- Analysis of panel data provides potentials and capabilities to address questions that would not have been possible using cross-section data.
- Specifically, with multiple observations per subject, we can understand behavior dynamic by observing the same subjects over time.
- We can also understand how these dynamics are related to other variables.
- Within-individual change is characterized in terms of some appropriate summary of the changes in the repeated measurements on each individual during the period of observation.

Characteristics of Panel Data and Implications

Two Key Characteristics:

- A common feature of repeated measurements on an individual is correlation, that is, knowledge of the value of the response on one occasion provides information about the likely value of the response on a future occasion.
- Another common feature of longitudinal data is heterogeneous variability, that is, the variance of the response changes over the duration of the study.

Characteristics of Panel Data and Implications

Consequences:

- These two features of longitudinal data violate the fundamental assumptions of independence and homogeneity of variance that are at the basis of many standard techniques (e.g., *t* test, ANOVA, and multiple linear regression).

Solution:

- To account for these features, statistical models for longitudinal data have two main components: a model for the covariance among repeated measures, coupled with a model for the mean response and its dependence on covariates.
  - Covariance means both the correlations among pairs of repeated measures on an individual and the variability of the responses on different occasions.
  - Failure to properly account for the covariance results in hypothesis tests and CIs that are invalid and may result in misleading inferences.

Berkeley SCHOOL OF INFORMATION

ANALYSIS OF PANEL DATA

An Introduction

datascience@berkeley

Using OLS Regression Model on Panel Data

OLS Regression

- The pros and cons of using OLS regression:
- **Pros**
  - Easy
  - Can apply a model that we've already learned
- **Cons:**
  - Estimates are unreliable.
  - Statistics are invalid as the key underlying assumptions of OLS models are violated.
  - Statistical inferences are incorrect.

OLS Regression – Let's Do It Anyway

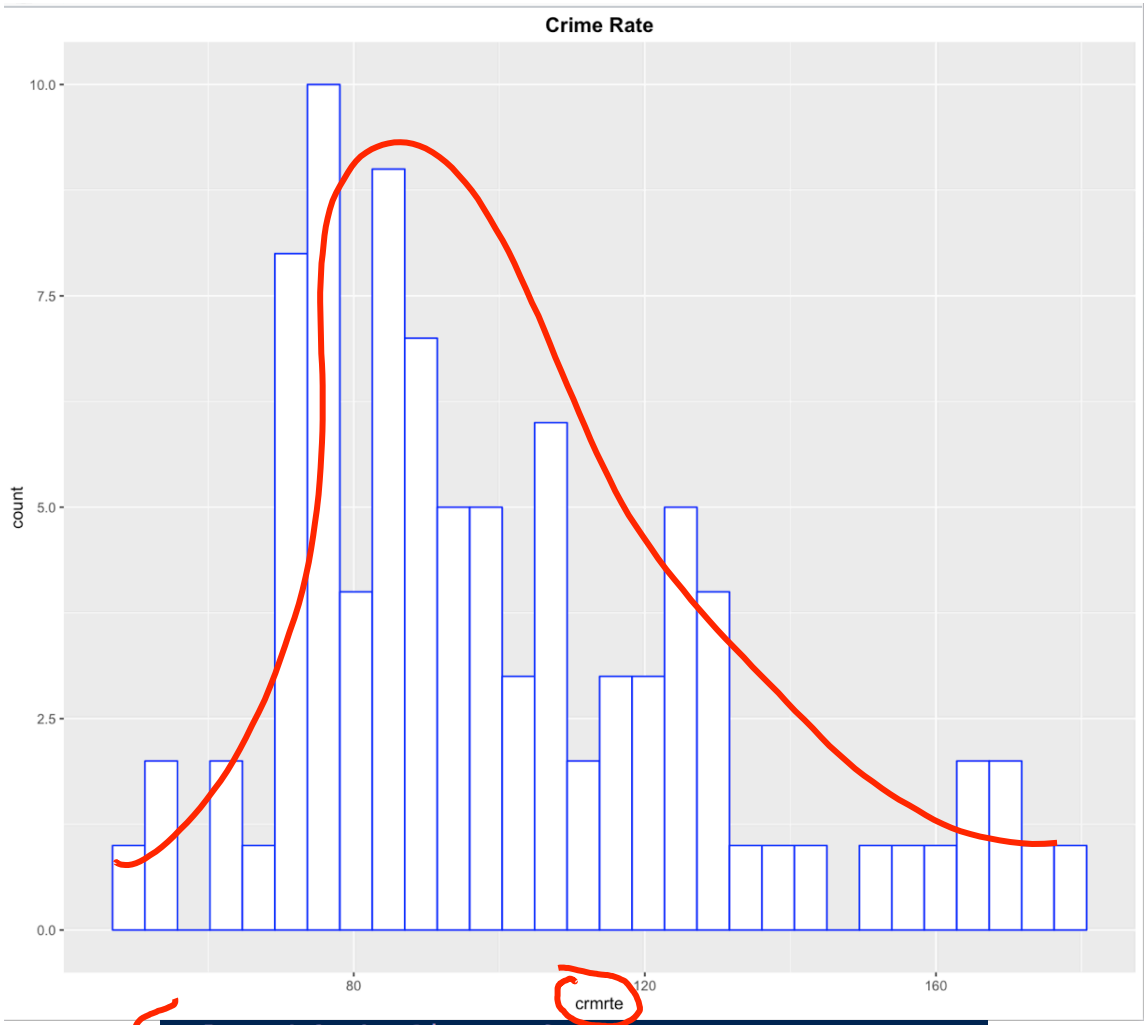
- Let's use a dataset that comes with Wooldridge's text: crime2.

```
'data.frame':   92 obs. of  34 variables:
 $ pop      : num  229528 246815 814054 933177 374974 ...
 $ crimes   : num  17136 17306 75654 83960 31352 ...
 $ unem     : num  8.2 3.7 8.1 5.4 9 ...
 $ officers: int  326 321 1621 1803 633 685 245 259 504 563 ...
 $ pcinc    : int  8532 12155 7551 11363 8343 11729 7592 10802
 $ west     : int  1 1 1 1 1 1 1 1 1 1 ...
 $ nrtheast: int  0 0 0 0 0 0 0 0 0 0 ...
 $ south    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ year     : int  82 87 82 87 82 87 82 87 82 87 ...
 $ area     : num  44.6 44.6 375 375 49.8 ...
 $ d87      : int  0 1 0 1 0 1 0 1 0 1 ...
 $ popden   : num  5146 5534 2171 2488 7530 ...
 $ crmrte   : num  74.7 70.1 92.9 90 83.6 ...
 $ offarea  : num  7.31 7.2 4.32 4.81 12.71 ...
 $ lawexpc  : num  851 2262 875 1070 1122 ...
 $ polpc    : num  1.42 1.3 1.99 1.93 1.69 ...
 $ lpop     : num  12.3 12.4 13.6 13.7 12.8 ...
 $ loffic   : num  5.79 5.77 7.39 7.5 6.45 ...
 $ lpcinc   : num  9.05 9.41 8.93 9.34 9.03 ...
 $ llawexpc: num  6.75 7.72 6.77 6.98 7.02 ...
 $ lpopden : num  8.55 8.62 7.68 7.82 8.93 ...
 $ lcrimes : num  9.75 9.76 11.23 11.34 10.35 ...
 $ larea    : num  3.8 3.8 5.93 5.93 3.91 ...
 $ lcrmrte : num  4.31 4.25 4.53 4.5 4.43 ...
 $ clcrimes: num  NA 0.00987 NA 0.10417 NA ...
 $ clpop    : num  NA 0.0726 NA 0.1366 NA ...
 $ clcrmrte: num  NA -0.0627 NA -0.0324 NA ...
 $ lpolpc   : num  0.351 0.263 0.689 0.659 0.524 ...
 $ clpolpc : num  NA -0.0881 NA -0.0302 NA ...
 $ cllawexp: num  NA 0.978 NA 0.201 NA ...
 $ cunem    : num  NA -4.5 NA -2.7 NA ...
 $ clpopden: num  NA 0.0726 NA 0.1366 NA ...
 $ lcrmrt_1: num  NA 4.31 NA 4.53 NA ...
 $ ccrmrte : num  NA -4.54 NA -2.96 NA ...
```

First Few Observations in the Dataset

```
> head(crime2)
     pop crimes unem officers pcinc west nrtheast south year  area d87   popden
1 229528  17136  8.2      326  8532    1        0     0   82  44.6   0 5146.368
2 246815  17306  3.7      321 12155    1        0     0   87  44.6   1 5533.969
3 814054  75654  8.1     1621  7551    1        0     0   82 375.0   0 2170.811
4 933177  83960  5.4     1803 11363    1        0     0   87 375.0   1 2488.472
5 374974  31352  9.0      633  8343    1        0     0   82  49.8   0 7529.599
6 406297  31364  5.9      685 11729    1        0     0   87  49.8   1 8158.574
    crmrte   offarea   lawexpc    polpc     lpop   loffic   lpcinc llawexpc  lpopden
1 74.65756  7.309417  850.8599 1.420306 12.34378 5.786897 9.051579 6.746247 8.546046
2 70.11729  7.197309 2262.4399 1.300569 12.41639 5.771441 9.405496 7.724199 8.618661
3 92.93487  4.322667  875.0800 1.991268 13.60978 7.390799 8.929436 6.774315 7.682856
4 89.97221  4.808000 1069.6400 1.932109 13.74635 7.497207 9.338118 6.975078 7.819424
5 83.61113 12.710844 1121.8999 1.688117 12.83461 6.450470 9.029179 7.022779 8.926597
6 77.19476 13.755020 1545.6000 1.685959 12.91484 6.529419 9.369820 7.343167 9.006824
    lcrimes    larea  lcrmrte     clcrimes      clpop    clcrmrte    lpolpc
1  9.748937 3.797734 4.312912           NA         NA          NA 0.3508723
2  9.758808 3.797734 4.250169 0.0098714828 0.07261372 -0.06274271 0.2628021
3 11.233926 5.926926 4.531899           NA         NA          NA 0.6887718
4 11.338096 5.926926 4.499501 0.1041698456 0.13656807 -0.03239822 0.6586123
5 10.353033 3.908015 4.426177           NA         NA          NA 0.5236138
6 10.353416 3.908015 4.346332 0.0003833771 0.08022785 -0.07984495 0.5223344
       clpolpc  cllawexp cunem   clpopden lcrmrt_1   ccrmrte
1           NA        NA    NA         NA       NA        NA
2 -0.088070214 0.9779520  -4.5 0.07261467 4.312912 -4.540268
3           NA        NA    NA         NA       NA        NA
4 -0.030159533 0.2007623  -2.7 0.13656807 4.531899 -2.962654
5           NA        NA    NA         NA       NA        NA
6 -0.001279354 0.3203883  -3.1 0.08022785 4.426177 -6.416374
```

Histogram of the Variable of Interest: Crimes




```
> length(crime2$crmrte)
[1] 92
> summary(crime2$crmrte)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  50.02   77.22   92.54  100.80  118.90  179.40
```

Berkeley SCHOOL OF INFORMATION

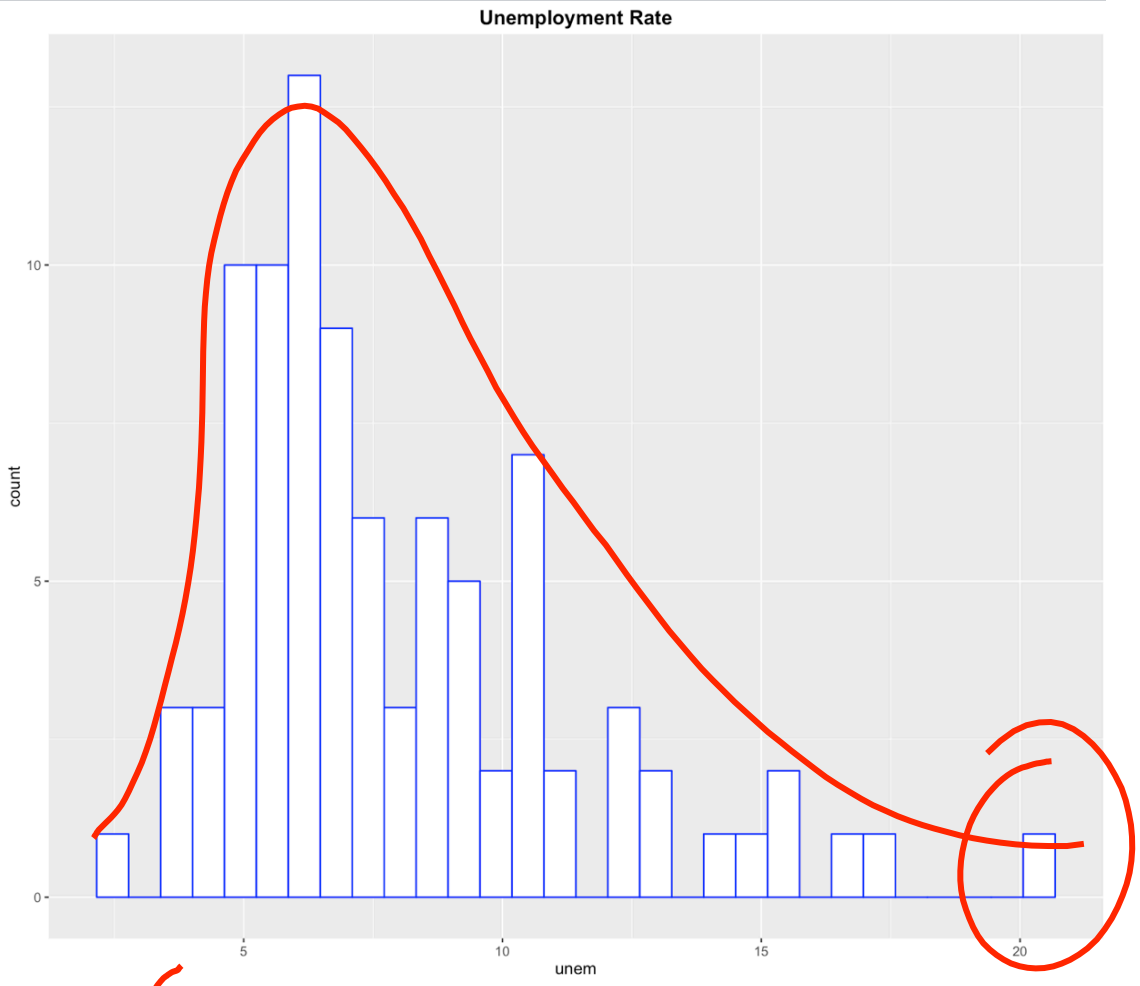
ANALYSIS OF PANEL DATA

An Introduction

datascience@berkeley

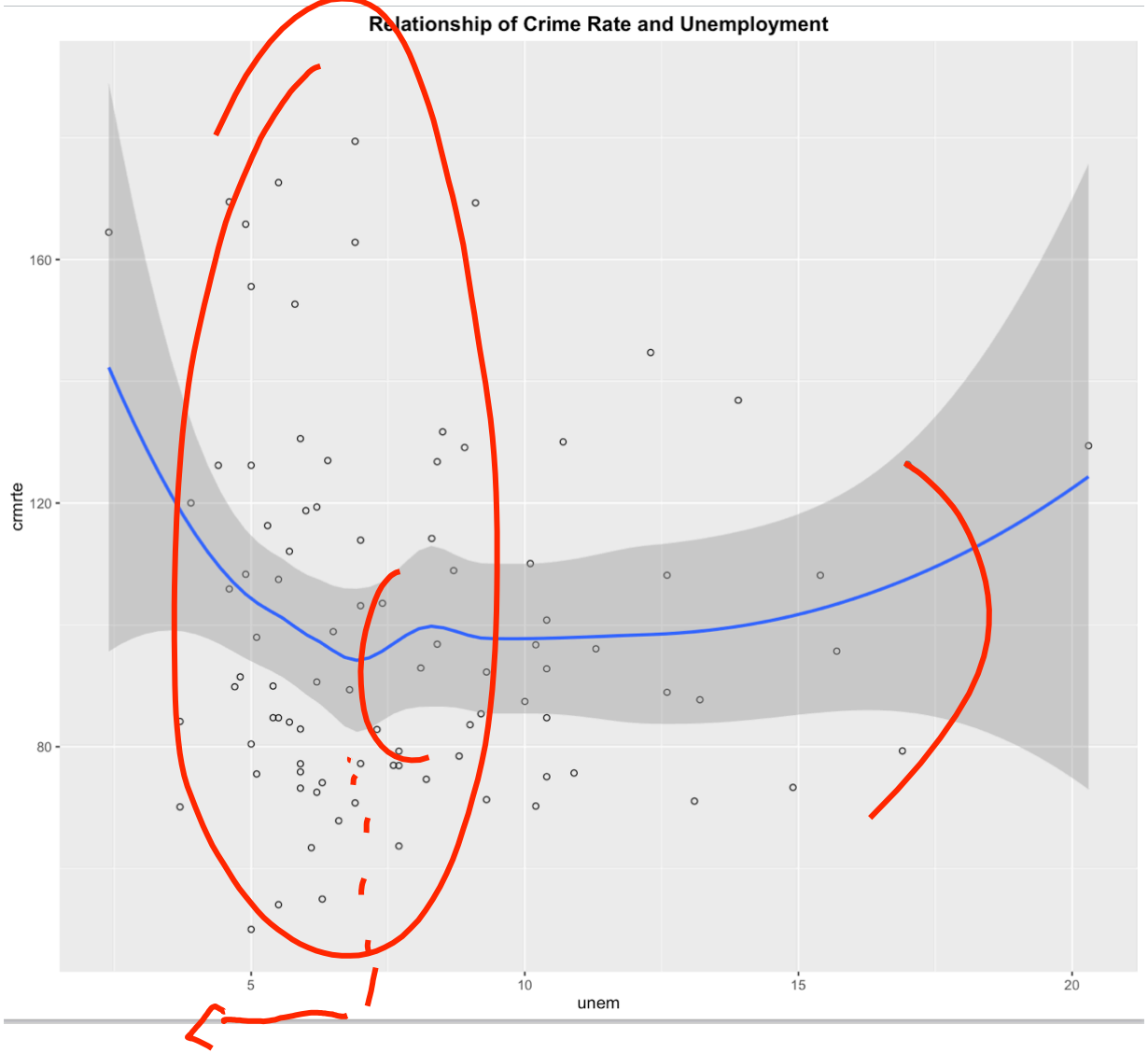
Using OLS Regression Model on Panel Data

Histogram of the Variable of Interest: unem




```
> length(crime2$unem)
[1] 92
> summary(crime2$unem)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.400   5.500   6.950   7.972   9.475  20.300
```

Relationship Between **crmrte** and **unem**

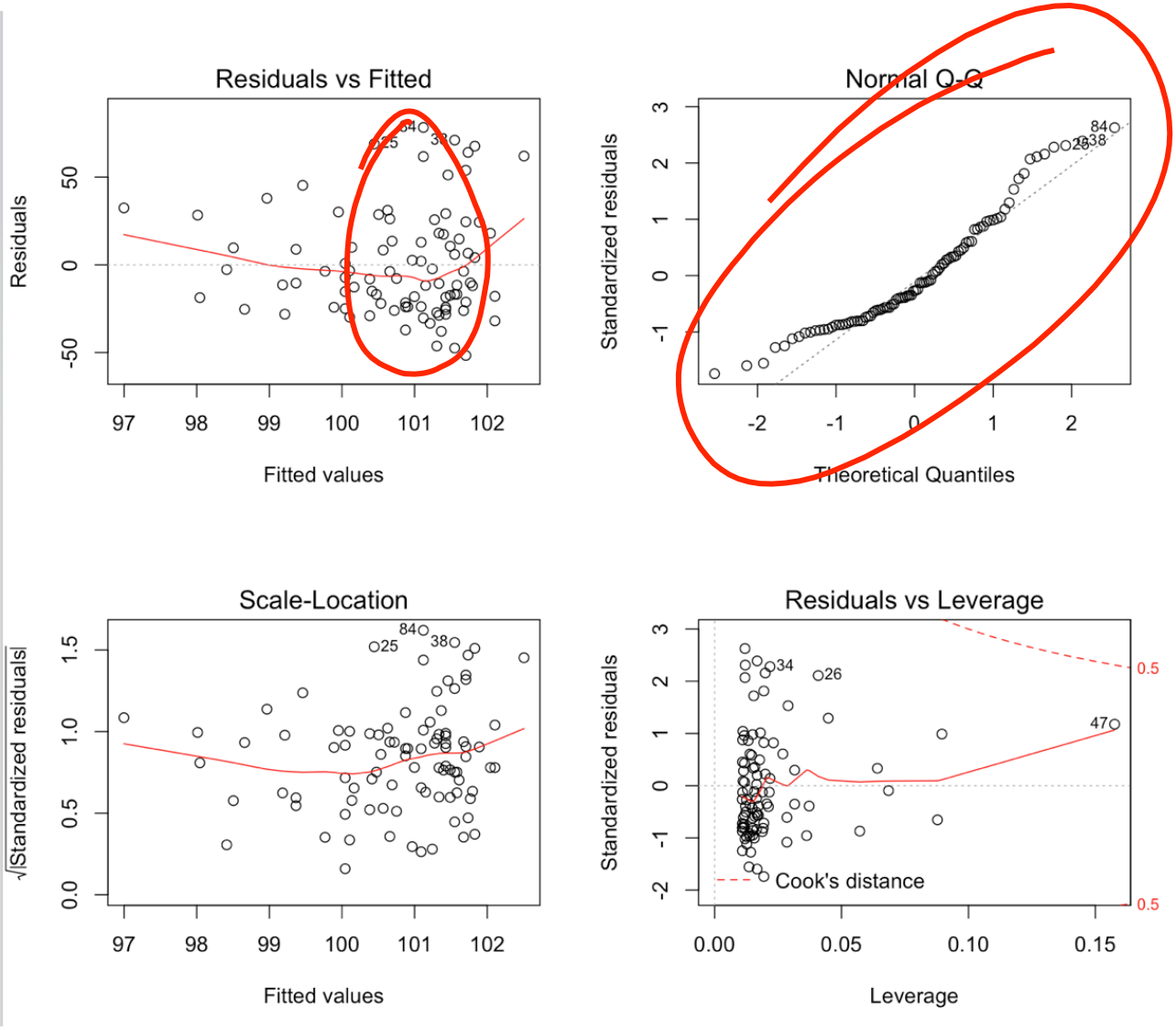


OLS Regression

In fact, let's run a simple OLS regression of **crmrte** on **unem** using all of the observations in the dataset. Note that I didn't do much EDA before building the model, but I just want to use this model to illustrate a point later.

```
> ols.fit1 <- lm(crmrte ~ unem, data=crime2)
> summary(ols.fit1)

Call:
lm(formula = crmrte ~ unem, data = crime2)

Residuals:
    Min      1Q  Median      3Q     Max
-51.686 -23.889  -7.961  17.522  78.297

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.2434     8.0587   12.81   <2e-16 ***
unem         -0.3077     0.9317   -0.33    0.742
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.99 on 90 degrees of freedom
Multiple R-squared:  0.00121,   Adjusted R-squared:  -0.009888
F-statistic: 0.109 on 1 and 90 DF,  p-value: 0.742
```

Residual Diagnostic Plots



Berkeley SCHOOL OF INFORMATION

ANALYSIS OF PANEL DATA

An Introduction

datascience@berkeley

Using OLS Regression Model on Panel Data

Structure of the Data

- We need to pay attention to the structure of the data.
- This (**crime2**) is a panel dataset in which each of the 46 cities was observed two times (1982 and 1987).
- We cannot simply treat these 46 "once repeated" observations as 92 independent observations.

OLS Regression Revisit

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

$$\underset{n \times (k+1)}{\mathbf{X}} \equiv \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}.$$

$\mathbf{X}$ is $n \times (k+1)$ and $\boldsymbol{\beta}$ is $(k+1) \times 1$, $\mathbf{X}\boldsymbol{\beta}$ is $n \times 1$.

Assumptions:

1. Linearity (in parameters)
2. $\mathbf{X}$ has rank $(k+1)$
3. $\mathrm{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}$
4. $\mathrm{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix.
5. $\mathbf{u} \sim \mathrm{Normal}(\mathbf{0}, \sigma^2\mathbf{I}_n)$

Potential Violations of Underlying Assumptions

- The independence of the independent and identically distribution (iid) is violated due to the repeated observations.
- In fact, the Durbin–Watson test confirms the violation of the independence assumption.

```
        Durbin-Watson test

data:  crmrte ~ unem
DW = 1.2074, p-value = 3.681e-05
alternative hypothesis: true autocorrelation is greater than 0
```

- When this assumption is violated, OLS standard errors and test statistics are not valid. Statistical inference becomes unreliable.

Berkeley SCHOOL OF INFORMATION

ANALYSIS OF PANEL DATA

An Introduction

datascience@berkeley

Exploratory Panel Data Analysis: A Two-Period Panel

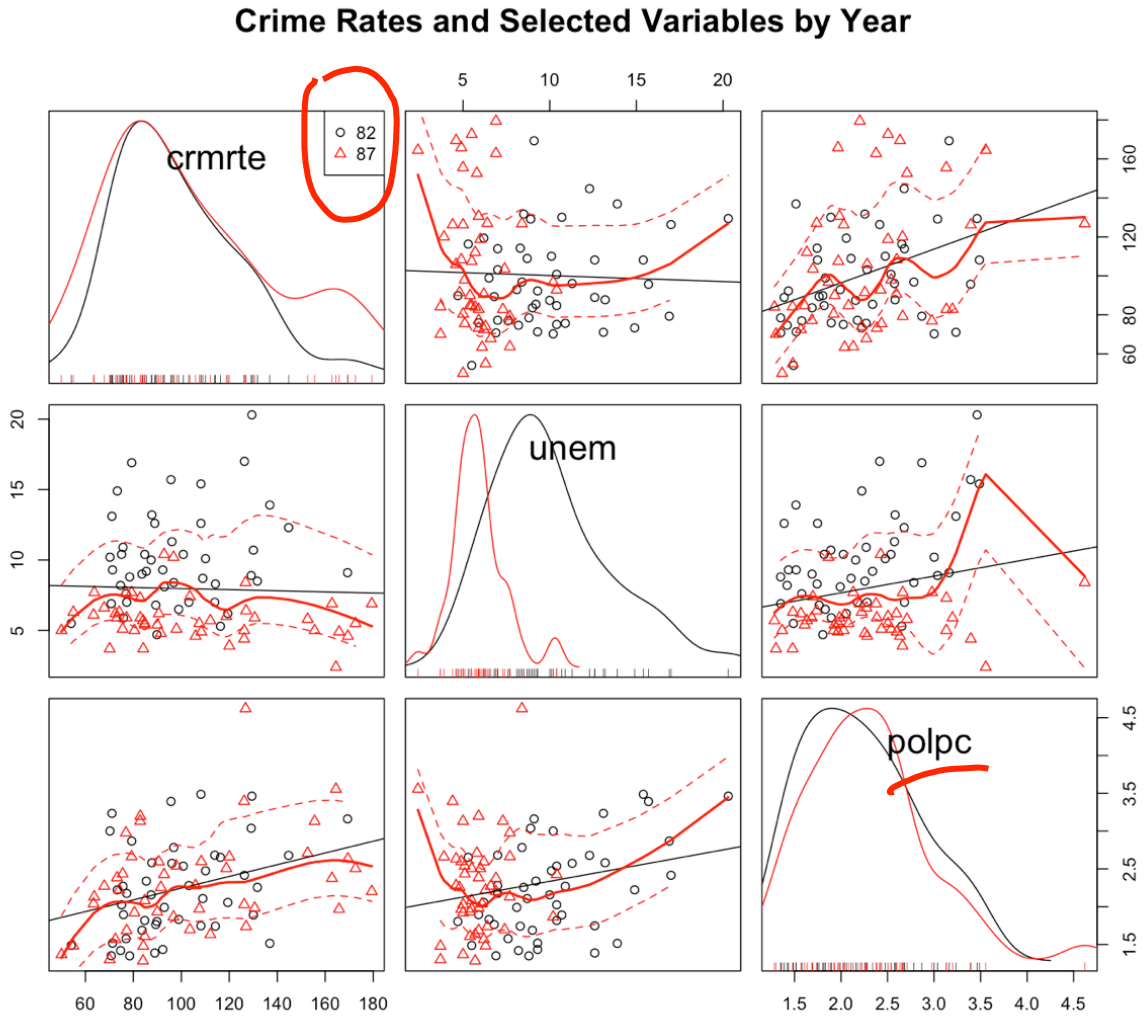
Exploring Panel Data

- The exploratory data analysis techniques studied in w203 and earlier in this course can be applied.
- However, with panel data, we first need to find out how many panels there are in the dataset.
- Then, not only do we need to explore patterns of each feature (as well as the combinations of the features) in each panel, we will have to explore the "dynamic" "temporal dependence" of the features.
- A subset of an example data (crime4) is shown here.

```
'data.frame':   630 obs. of  59 variables:
 $ county  : int  1 1 1 1 1 1 1 3 3 3 ...
 $ year    : int  81 82 83 84 85 86 87 81 82 83 ...
 $ crmrte  : num  0.0399 0.0383 0.0303 0.0347 0.0366 ...
 $ prbarr  : num  0.29 0.338 0.33 0.363 0.325 ...
 $ prbconv : num  0.402 0.433 0.526 0.605 0.579 ...
 $ prbpris : num  0.472 0.507 0.48 0.52 0.497 ...
 $ avgsen  : num  5.61 5.59 5.8 6.89 6.55 ...
 $ polpc   : num  0.00179 0.00177 0.00184 0.00189 0.00192 ...
 $ density : num  2.31 2.33 2.34 2.35 2.36 ...
 $ taxpc   : num  25.7 24.9 26.5 26.8 28.1 ...
 $ west    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ central : int  1 1 1 1 1 1 1 1 1 1 ...
 $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pctmin80: num  20.2 20.2 20.2 20.2 20.2 ...
 $ wcon    : num  206 213 220 223 244 ...
 $ wtuc    : num  334 369 1395 399 359 ...
 $ wtrd    : num  182 190 197 201 207 ...
 $ wfir    : num  272 301 310 350 383 ...
 $ wser    : num  216 232 240 252 261 ...
 $ wmfg    : num  229 240 270 282 299 ...
 $ wfed    : num  409 420 439 459 490 ...
 $ wsta    : num  236 254 250 262 281 ...
 $ wloc    : num  231 237 249 264 289 ...
 $ mix     : num  0.0999 0.103 0.0807 0.0785 0.0932 ...
 $ pctymle : num  0.0877 0.0864 0.0851 0.0838 0.0823 ...
 $ d82     : int  0 1 0 0 0 0 0 0 1 0 ...
 $ d83     : int  0 0 1 0 0 0 0 0 0 1 ...
 $ d84     : int  0 0 0 1 0 0 0 0 0 0 ...
 $ d85     : int  0 0 0 0 1 0 0 0 0 0 ...
 $ d86     : int  0 0 0 0 0 1 0 0 0 0 ...
 $ d87     : int  0 0 0 0 0 0 1 0 0 0 ...
```

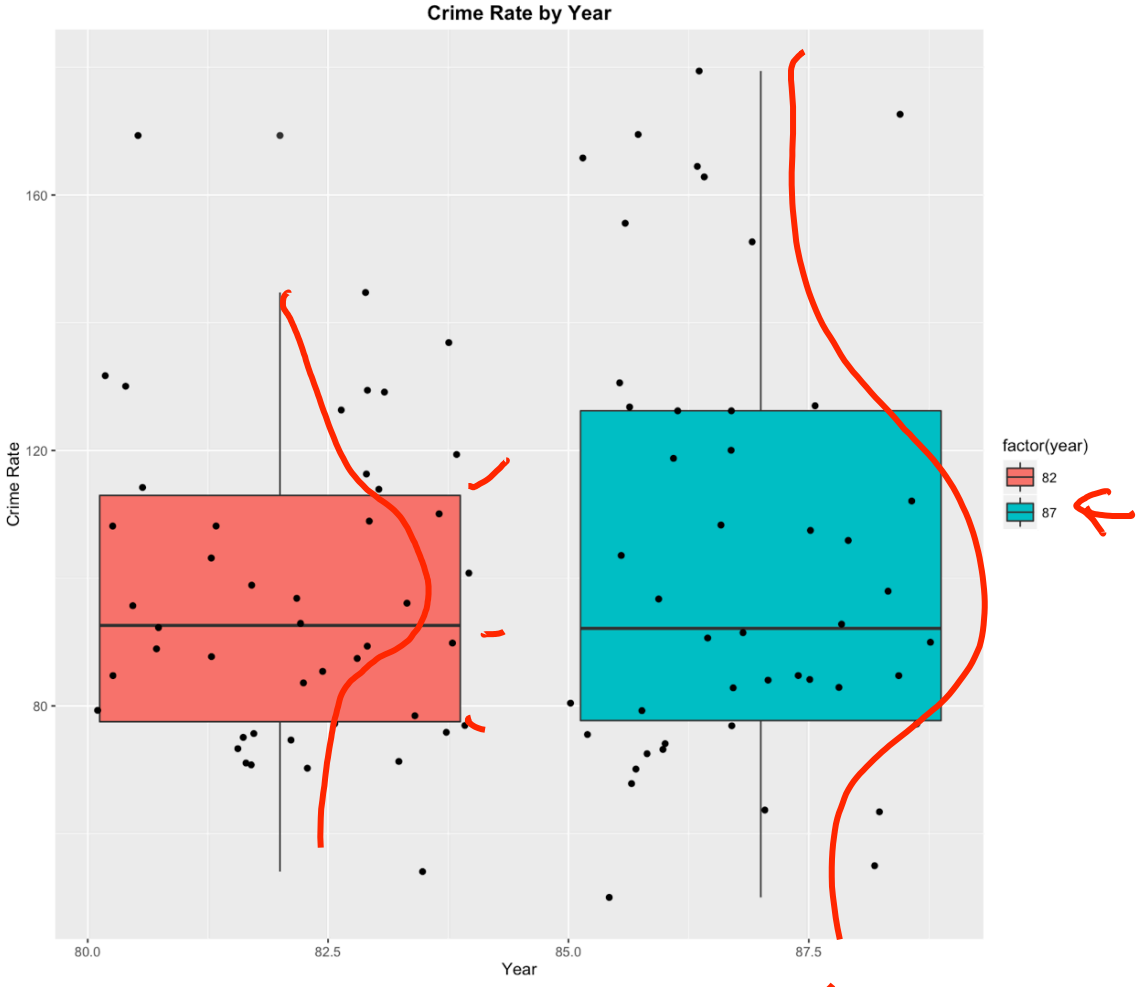
```
> table(crime4$year)

81 82 83 84 85 86 87
90 90 90 90 90 90 90
```

Distributions and Dependence Over Time



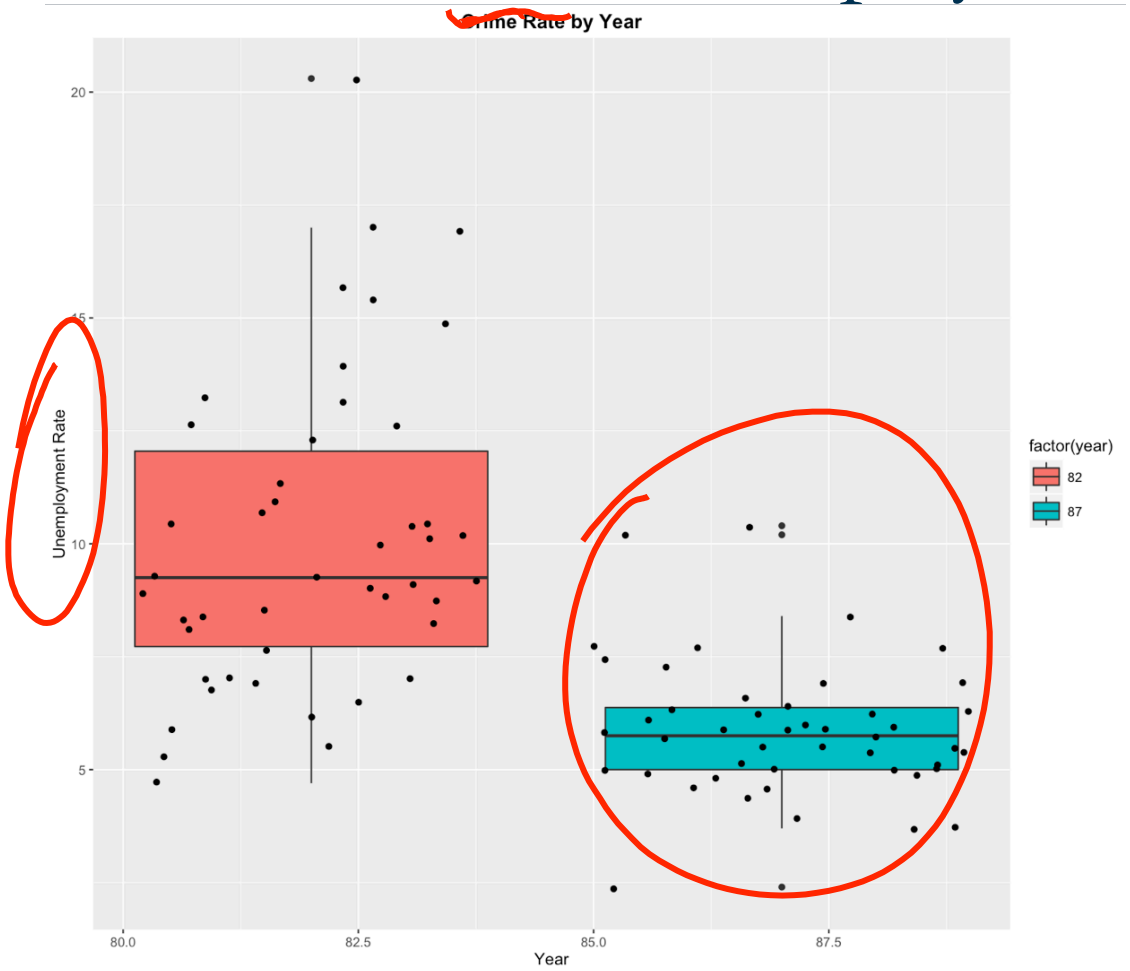
- Recognize that relationship between variable of interest and predictors may change over time.
- The distribution of each of the variables may also change over time.

Distribution of Crime Rate Over Time



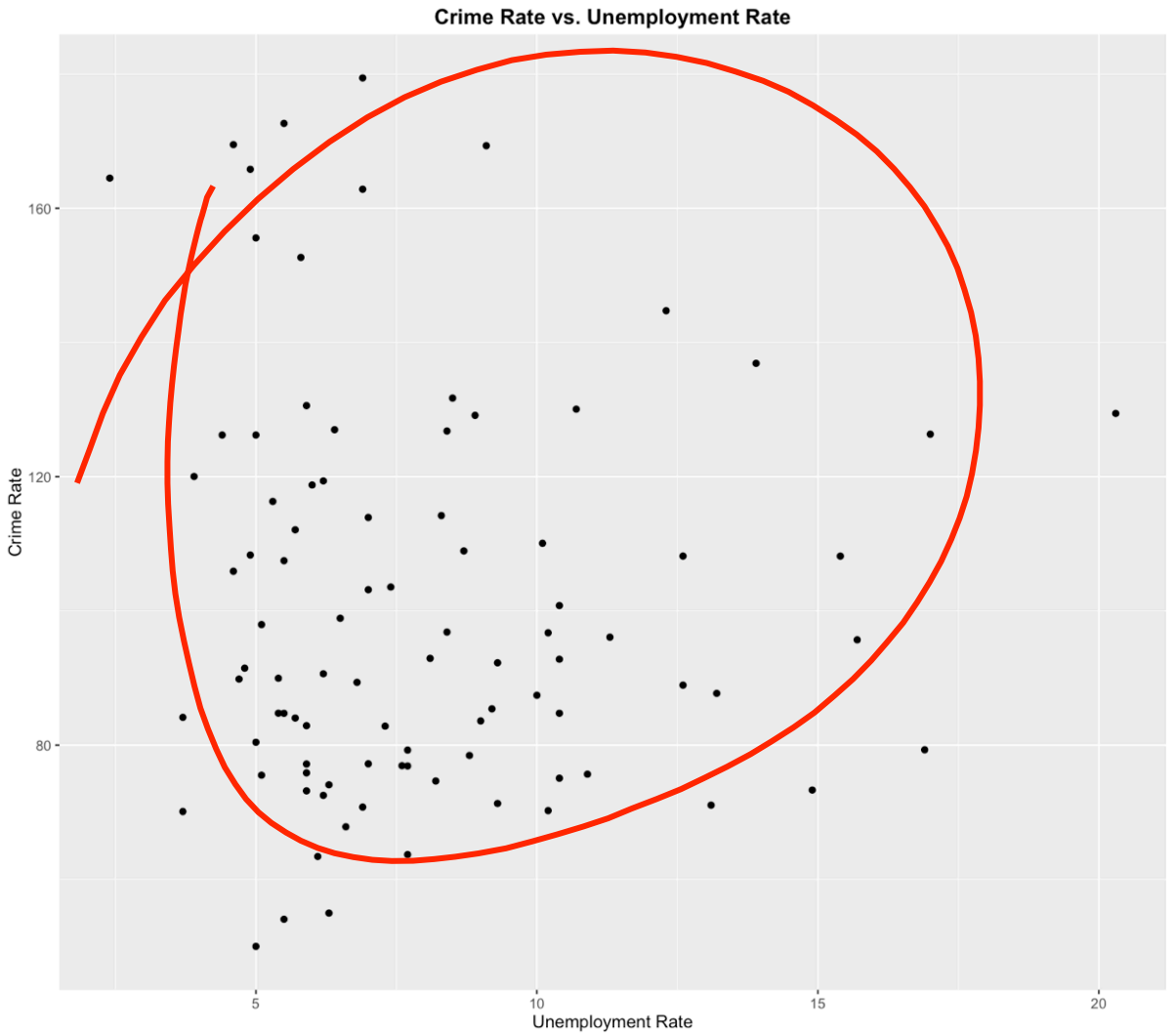

```
summary(crime2$crmrte[crime2$year==82])
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
54.06   77.54   92.61   97.71  113.00  169.30
summary(crime2$crmrte[crime2$year==87])
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
50.02   77.71   92.14  103.90  126.20  179.40
```

Distribution of Unemployment Rate Over Time




```
> summary(crime2$unem[crime2$year==82])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.700   7.725   9.250  10.050  12.050  20.300
> summary(crime2$unem[crime2$year==87])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.400   5.000   5.750   5.889   6.375  10.400
```

Relationship Between Crime Rate and Unemployment Rate



Relationship Between Crime Rate and Unemployment Rate Changed Over Time

Notice the change in relationship between crime rate and unemployment rate over time.

Naïve OLS Regression 1 : Using Only the 1982 Panel

```
> ols.fit1<-lm(crmrte ~ unem, data=crimes.82)
> summary(ols.fit1)

Call:
lm(formula = crmrte ~ unem, data = crimes.82)

Residuals:
    Min      1Q  Median      3Q     Max
-37.693 -17.292  -3.671  16.994  72.854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   84.569     10.904   7.756 9.07e-10 ***
unem           1.307      1.027   1.272     0.21
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.74 on 44 degrees of freedom
Multiple R-squared:  0.03549,   Adjusted R-squared:  0.01357
F-statistic: 1.619 on 1 and 44 DF,  p-value: 0.2099
```

Berkeley SCHOOL OF INFORMATION

ANALYSIS OF PANEL DATA

An Introduction

datascience@berkeley

Exploratory Panel Data Analysis: A Two-Period Panel

Naïve OLS Regression 2: Using Only the 1987 Panel

```
> ols.fit2<-lm(crmrte ~ unem, data=crimes.87)
> summary(ols.fit2)

Call:
lm(formula = crmrte ~ unem, data = crimes.87)

Residuals:
   Min     1Q Median     3Q    Max
-57.55 -27.01 -10.56  18.01  79.75

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  128.378     20.757   6.185  1.8e-07 ***
unem          -4.161      3.416  -1.218     0.23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.6 on 44 degrees of freedom
Multiple R-squared:  0.03262,   Adjusted R-squared:  0.01063
F-statistic: 1.483 on 1 and 44 DF,  p-value: 0.2297
```

$$\hat{crmrte} = 123.38 - 4.16 unem$$

where $n = 46, R^2 = 0.033$

Berkeley SCHOOL OF INFORMATION

ANALYSIS OF PANEL DATA

An Introduction

datascience@berkeley

Unobserved Effect Models and Pooled OLS and First-Difference Models

Panel Data and Unobserved Effect

- In the previous example, we have a cross section of cities observed in two different years.
- In general, the observational units can be individuals, companies, schools, countries, and so on.
- However, the two naïve OLS regression models estimated above are likely suffer from omitted variable problems.
- One can argue to include more observable explanatory variables in the regression model, such as education level, age distribution, gender distribution, and so on.
- There are a few alternative ways to utilize information available in a panel dataset in order to both deal with unobserved variables and capture the dynamic that would not have been possible using cross-section data.
- There are two types of unobserved variable (or unobserved effect): (1)those that are time invariant, and (2) those that are time varying.
- Note that this and the next two lectures are conceptually more abstract, and the mathematical notations used are more involved. Students are reminded to read the assigned chapters.

Simple Formulation With Unobserved Effect

Let i denote the cross-sectional unit and t the time period. A very simple formulation that includes a single observed explanatory variable and an unobserved variable is

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + \epsilon_{it}$$

where $t = 1, 2$

$d2_t$ is a dummy variable equal to 0 when $t = 1$ and 1 when $t = 2$.

ϵ_{it} is the idiosyncratic error, which, in the current context, is time-varying.

a_i captures all unobserved, time-invariant variables that affect y_{it}. In econometrics, it is often called **unobserved effect**, **fixed effect**, or even **unobserved heterogenity**. As such, the model above can be called an **unobserved effect model** or a **fixed effect model**.

Berkeley SCHOOL OF INFORMATION

ANALYSIS OF PANEL DATA

An Introduction

datascience@berkeley

Unobserved Effect Models
Pooled OLS
First-Difference Models

Unobserved City Effect in Our Example

With this set up, a simple unobserved effect model for city crime rates for 1982 and 1987 is

$$crmrte_{it} = \beta_0 + \delta_0 d87_t + \beta_1 unem_{it} + a_i + \epsilon_{it}$$

where $d87$ is an indicator variable for year 1987, and i denotes cities.

In this example, a_i is called an **unobserved city effect** or **city fixed effect**.

Any city-specific features, such as geographical features, that do not change over *the observed time period* and are not observed (by the analysts) are included in a_i. We have to emphasize *constant only observed time period* because every geographical features are not constant forever. Also pay attention to the subscripts being use: a_i only varies across cross-sectional units (and not over time) but constant within each of the cross-sectional unit.

Pooled OLS Applied to the Crime Rate Example

One method is to "*pool*" the two years and use **OLS**. The major drawback of this approach is that the pooled OLS requires that the observed effect a_i and the observed explanatory variable, x_{it}, be *uncorrelated* in order to produce a *consistent* estimator for β_1.

Writing the model using a *composite error* form, we have

$$crmrte_{it} = \beta_0 + \delta_0 d87_t + \beta_1 unem_{it} + \mu_{it}$$

where $\mu_{it} = a_i + \epsilon_{it}$.

- OLS requires that μ_{it} be uncorrelated with x_{it}. While it may be a reasonable assumption in cross-sectional regression, this assumption is not likely to hold because the same cross-sectional units are observed multiple times.
- Even if ϵ_{it} is uncorrelated, the pooled OLS is likely to be *biased* and *inconsistent* when a_i and x_{it} are correlated. This kind of bias is called *heterogeneity bias*.
- The term *heterogeneity bias* here is referred to the bias is really caused by omitting individual-specific, time-invariant variables.

Pooled OLS Applied to the Crime Rate Example

As you can see in the example above, pooled OLS without even the time indicator variable $d87$ produces unreasonable results in addition to violating the correlation assumption and potentially suffering from omitted variables.

Let's try another pooled OLS model:

```
> pooled.ols.fit <- lm(crmrte ~ d87+unem, data=crime2)
> summary(pooled.ols.fit)

Call:
lm(formula = crmrte ~ d87 + unem, data = crime2)

Residuals:
    Min      1Q  Median      3Q     Max
-53.474 -21.794  -6.266  18.297  75.113

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  93.4202    12.7395   7.333 9.92e-11 ***
d87           7.9404     7.9753   0.996    0.322
unem          0.4265     1.1883   0.359    0.720
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.99 on 89 degrees of freedom
Multiple R-squared:  0.01221,   Adjusted R-squared:  -0.009986
F-statistic: 0.5501 on 2 and 89 DF,  p-value: 0.5788
```

Estimated Model Using Pooled OLS

$$\widehat{crmrte} = 93.42 + 7.94 d87_t + 0.427 unem$$

where $n = 92$ and $R^2 = 0.012$

We drop the subscripts when reporting the estimated model.

- Although the estimated effect of unemployment rate on crime rate directionally makes sense, it is both economically and statistically insignificant.
- The model also does not explain the crime rate well.
- Most importantly, the standard errors and test statistics in this model are incorrect due to the serial correlation caused by the repeated observations. This is a point mentioned before.

The First-Differencing Approach

- Panel data statistical models face this problem directly without making unreasonable assumption regarding the absence of correlation between the individual heterogeneity and the explanatory variables.
- In fact, this kind of models allow for the unobserved effect, a_i, to be correlated with explantory variables.
- In our example, we would like to allow for correlation between the unobserved (to the data scientists) city variables that affect crime rate and the observed explanatory variables, such as unemployment rate.
- For unobserved effect, a_i, that is individual-specific and remain constant over the observed time period, we can use differencing to eliminate the unobservables while estimating the effect of interest.

The First-Differencing Approach

For example,

$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2} \quad (t = 2)$$
$$y_{i1} = \beta_0 + \beta_1 x_{i1} + a_i + u_{i1} \quad (t = 1).$$

Substracting the second equation from the first, we get

$$(y_{i2} - y_{i1}) = \delta_0 + \beta_1 (x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$$

or

$\triangle y_i = \delta_0 + \beta_1 \triangle x_i + \triangle u_i$ where $\triangle$ denotes the change from $t = 1$ to $t = 2$.

Note that in the "change" equation, the unobserved effect is "*differenced*" away.

The First-Differencing Approach

- The "first-difference" equation is very simple in that it is a cross-sectional equation in which each of the variable is differenced over two consecutive time periods.

- Importantly, we can estimate this model and conduct inference using the estimated models by OLS regression techniques, provided that the underlying assumptions are satisfied.

- Specifically, this model requires that $\triangle u_i$ is uncorrelated with $\triangle x_i$.

- This assumption would hold if the error term, u_{it}, is uncorrelated with the explanatory variables in *both* time periods. This is the version of **strict exogenity** assumption.

The First-Differencing Approach

$$\widehat{\Delta crmrte} = 15.40 + 2.22\ \Delta unem$$
$$(4.70) \quad (.88)$$
$$n = 46,\ R^2 = .127,$$

- The differencing to eliminate time-invariant unobservable effects has a substantial inference on the results (relative to the OLS models estimated above). Each 1% change in unemployment rate is associated with an average of a 2.2 increase in crime rate, measured by the number of crimes per $1,000$ residents.

The First-Differencing Approach

- Differencing in this case also makes intuitive sense because instead of estimating a cross-sectional relationship, which possibly suffers from omitted variables, as it models directly how changes in the explanatory variables over time (in this toy example, unemployement rate) affects the change in y (in this case, crime rate) over the same time period.

- One has to remember that this approach will not work for explanatory variables that are "" over the observable time period of interest as they will be differenced out along with the time invariant unobservables.

constant

Berkeley SCHOOL OF INFORMATION

ANALYSIS OF PANEL DATA

An Introduction

datascience@berkeley

Distributed Lag Models

Distributed Lag of Clear-Up Rate on Crime Rate

- Eide (1994) uses a panel data from police districts in Norway to estimate a *distributed lag* model for crime rate.
- In this example, we use only a single explanatory variable, but in a distributed lag framework.
- The explanatory varible is "*clear-up percentage (clrpc)*", the percentage of crime that leads to conviction.
- The crime rate data are collected on year 1972 to 1978. In this example, we follow Eide and use two lags. The intuition is that the past clear-up rate may have a deterrent effect on current crime rate. From a policy perspective, one could think of it as "current clear-up rate may have a deterrent effect on future crime rate."

$$log(crime_{it}) = \beta_0 + \delta_0 d78_t + \beta_1 clrpc_{i,t-1} + \beta_2 clrpc_{i,t-2} + a_i + u_{it}$$

The Estimated Model

$$\widehat{\Delta \log(crime)} = .086 - .0040\ \Delta clrprc_{-1} - .0132\ \Delta clrprc_{-2}$$
$$(.064) \quad (.0047) \qquad\qquad (.0052)$$
$$n = 53,\ R^2 = .193,\ \bar{R}^2 = .161.$$

- The second lag is statistically significant and is negative, implying that a higher clear-up rate two years ago would deter crime rate in the current year. Specifically, a 10% increase in $clrpc$ two years ago would lead to an estimated 13.2% decline in crime rate in the current year (in the time periods in the dataset).
- It is very important to remember that this model is estimated using data from 1972 to 1978. The country might be very different now, and the same estimated effect may not be applicable to the current environement. As a data scientist, you should always keep in mind the purpose of the building and estimating a model.

Berkeley SCHOOL OF INFORMATION