

## Lectures 8: Model Assumptions, Part 2

- Model assumptions for Linear Regression (**review**)
- What can go wrong
- An Example in which some assumptions are not satisfied
  - House prices in Zip 30062 (A closer look!)
- “Repair”
  - **Removal (or not) of Outliers**
  - Transformations (**partly done in Lecture 7**)
    - Polynomial regression is also a possibility. To be discussed later.
  - Addressing Influential Points

# Ordinary Linear Model Assumptions (review)

- Properties of errors under ideal model:
  - $\mu_{Y|x} = \beta_0 + \beta_1 x$  for all  $x$ .
  - $Y_i = \beta_0 + \beta_1 x_i + e_i$  for all  $x_i$
  - All  $e_i$  have the same variance,  $\sigma_e^2$  [“Homoscedasticity”]
  - The distribution of  $e_i|x_i$  is normal. [Also  $E(e_i|x_i) = 0$ ]
  - $e_1, \dots, e_n$  are independent.
- Equivalent description:

For each  $x_i$  the corresponding  $Y_i$  has a normal distribution with mean  $\beta_0 + \beta_1 x_i$ , a linear function of  $x$  and constant variance. Also  $Y_1, \dots, Y_n$  are independent.

## Things to look for in the data (review)

These things suggest deviations from the assumptions

### Order of Inspection

1. Outliers (Covered in Lecture 7)

2. Non-linearity (Partly covered in Lecture 7)

– A Curved overall pattern as a function of  $x$ .

3. Heteroscedasticity

– Changes in the *spread* of the data for varying  $x$

4. Non-normality of the residuals

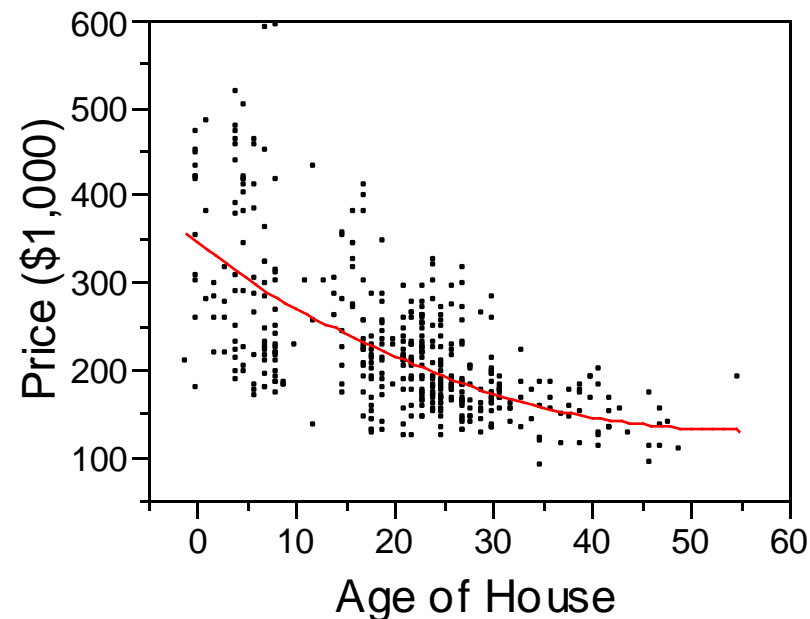
5. Lack of Independence

– This is fairly rare except in time-series data, *which will be studied later*

6. Influential Points: These are not a “deviation from the assumptions”, but they do require caution in interpreting the analysis

## Example: House Prices in Zip 30062 (*Review from Lecture 7*)

We look at House Prices (for sales in 2003) as a function of the age of the house.



The smooth curve through the data is called a “*smoothing spline*”. The details of its construction are not important (for now). It follows the general pattern of the average of  $Y$  at each given  $x$ .

## 2. Non-linearity (review from Lecture 7)

- This is a curved general pattern of the means of  $Y$  given  $x$ .
- Our data follows such a curved, non-linear pattern.
  - This is evident from the curved **spline** on the plot
  - It's pretty easy to see even without that **red** curve
- Non-linear strongly suggests that the general model **is not valid**.
- Hence, the least squares line, and associated tests and CI<sup>s</sup>, are **not** faithful representations of reality
  - *If the plot isn't too non-linear the LS analysis may still be somewhat useful.*
- There are various things that can be done to produce a valid analysis. (*See following slides*)

## 2. Non-linearity (review)

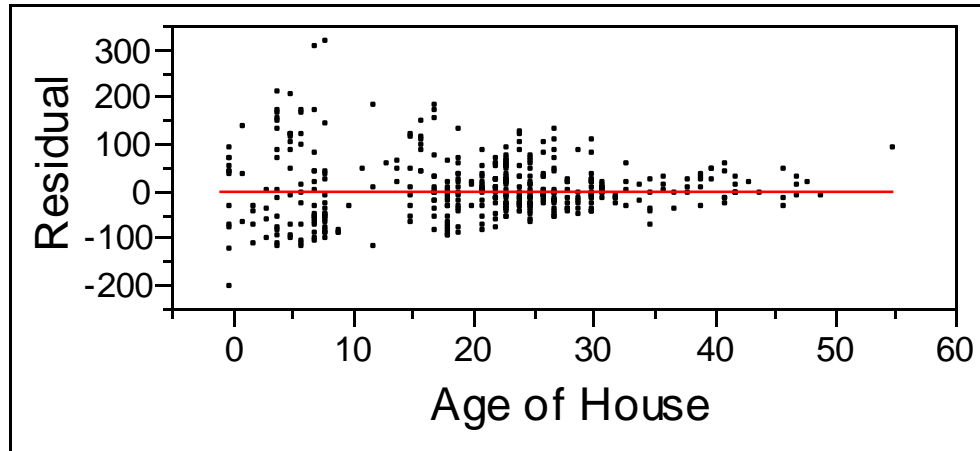
### Transformations

- Transformations of the variables  $x$  and/or  $Y$  may be useful.
- Transformations of  $Y$  will affect the scedasticity (variance) of  $Y$  given  $x$ .
  - They are best suited for use when heteroscedasticity is also present.
- In seemingly homoscedastic situations they should only be used with extreme caution to check for heteroscedasticity *after* use of the transformation
- Transformations of the  $x$  variable
  - Covered in Lecture 7

### 3. Heteroscedasticity

- This refers to situations in which the population model's *error variances for given  $x$*  are not equal as  $x$  changes
- Look for *vertical spreads* in the scatterplot (or in the plot of residuals) that gradually  $\nearrow$  as  $x \nearrow$  or that gradually  $\searrow$  as  $x \nearrow$ .
- *Other patterns of vertical spread can occasionally occur and be of interest.*
- We want fairly convincing evidence that the population model is not homoscedastic
  - Do not be overly affected by a few scattered points.
- Look at our **scatterplot**. In spite of the non-linearity confusing the view, you should still see the heteroscedasticity.
- The heteroscedasticity is much more evident in the **residual plot** about the SqRt fit; This is typically the better plot to look at in order to spot heteroscedasticity. Here is that plot --

## Residual Plot for the residuals from the **SqRt** fit



- Such a plot shows the values of the residuals as a function of  $x$ .
  - The “Residuals” are the values of  $Y_i - \text{Fit}(x_i)$
- You should see that the vertical spread at each  $x$  *generally*  $\searrow$  as  $x \nearrow$ . This is heteroscedasticity in the data.

### 3. Heteroscedasticity (cont)

#### Transformations of $Y$

- These can often successfully address apparent heteroscedasticity
- They can also address non-linearity – or they can take a linear situation and make it non-linear!
- So they must be used with care and caution
  - Look at the transformed model &  
Determine whether it seems to satisfactorily satisfy **both** linearity and homoscedasticity
  - Our data is an example where this happens
- It's possible that transforming  $Y$  may 'solve' heteroscedasticity, but not yield linearity
  - In that case a further transformation of  $x$  may help

*We'll later see an example of this sort*

## Transformations of $Y$ (cont)

<i>Group A</i>	<i>Group B</i>
$y'(y) = \sqrt{y}$	$y'(y) = y^2$
$y'(y) = \ln(y)$ or $Log_{10}(y)$	$y'(y) = e^y$
$y'(y) = 1/y$	

- Effect on “Scedasticity”:

**Group A** helps cure:

‘vertical spread larger when **mean of  $Y|x$**  is larger’

**Group B** helps cure:

‘vertical spread smaller when **mean of  $Y|x$**  is larger’

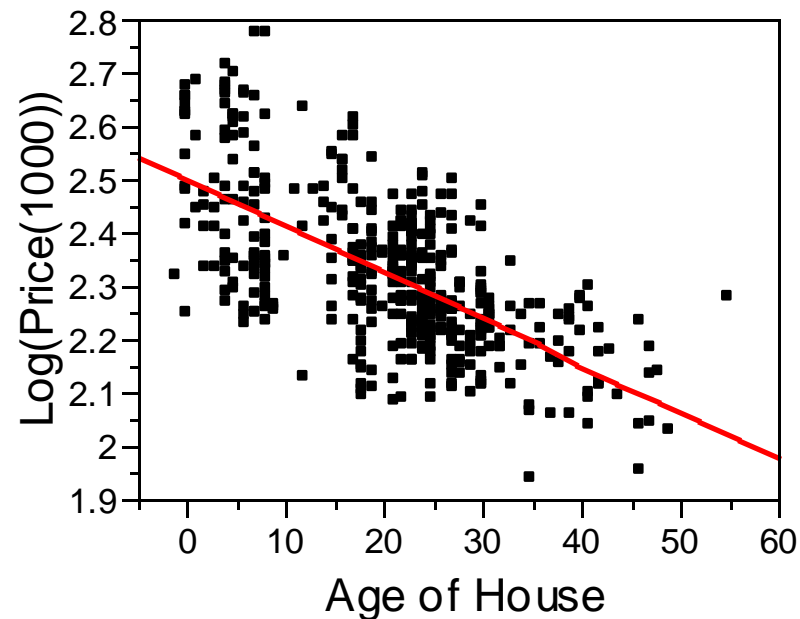
## House Price Data for Price vs Age of House (cont)

- Our data has ‘vertical spread larger when mean of  $Y|x$  is larger’
- It is also monotone (decreasing) and convex *ie*, the second derivative is positive
- HENCE, we should try a **Group A** transformation of  $Y$ .
- The log transformations seem most appropriate  
*(more about this after we look at the plot)*
- Either of the transformations  $y'(y) = \ln(y)$  or  $Log_{10}(y)$  have exactly the same qualitative effect – so either can be used
- The transformation  $y'(y) = Log_{10}(y)$  may be a little easier to interpret, since  $Log_{10}(y)$  is easier to understand
- Here’s the result from using  $y'(y) = Log_{10}(y)$
- Ways to do this in JMP will be described in Lecture

– 2 ways for now; one more way later (esp for CIs)

## House Price Data Transformed by $y'(y) = \text{Log}_{10}(y)$

NOTE: This scatterplot has  $y'(y) = \text{Log}_{10}(y)$  on the vertical axis, not  $y$  as did the previous plots

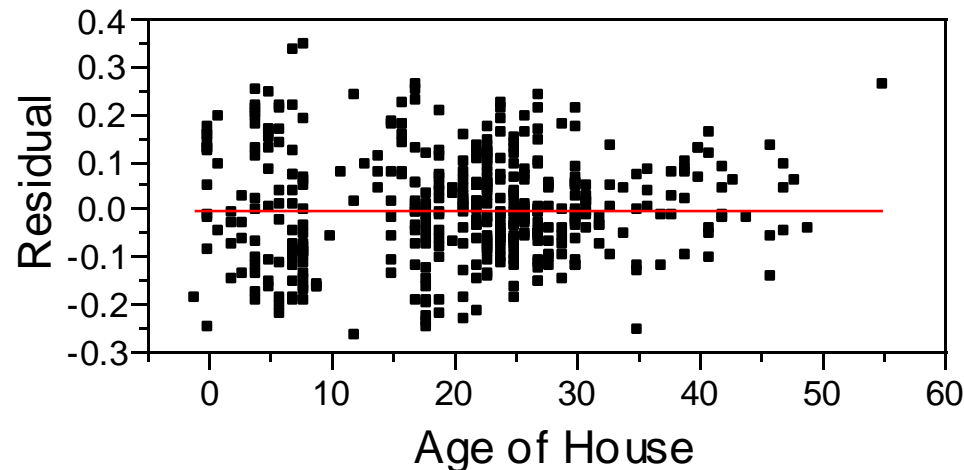


**This plot is satisfactorily linear and homoscedastic.**

Residual Plot for the analysis that uses  $y'(y) = \text{Log}_{10}(y)$

The residual plot for the previous analysis can help visually check linearity and homoscedasticity. Here it is

(remember,  $\text{Log}_{10}Y$  is the original variable for the transformed regression)



**This plot shows satisfactory linearity and homoscedasticity.**

- We could have guessed that this would be so, because the original  $Y$  variable (HousePrice) is a monetary quantity

## Special Role of Log (or ln) Transformations for financial data

- Log transformations very often appropriate for monetary data, *eg*,
  - Wages, income, earnings
  - Yields on investments and/or Profits
  - Costs of goods, *etc*
- That's how money works!
  - For an item that costs \$20 a \$2 price reduction to \$18 matters
  - BUT for an item that costs \$20,000 a reduction to \$19,998 doesn't (*or, shouldn't*) matter. **\$2 is very little in relation to \$20,000!**
  - HOWEVER for the \$20,000 item a reduction to \$18,000 matters
- What matters is the % reduction or increase -

In the above examples it's a **20%** reduction that matters

(cont)

## Log Transformations (cont)

- A 20% change in price,  $Y$ , involves *multiplying*  $Y$  by 1.2
- But it involves *adding*  $\text{Log}(1.2) = .079$  to  $\text{Log}(Y)$
- Additive changes do not affect standard deviations
- SO we can expect the values of  $\text{Log}(Y)$  at various values of  $x$  to be homoscedastic; and to show change only in their mean and not in their vertical spread (= standard deviations at given  $x$ )

## Log Transformations (cont)

- With this in mind, let's take a closer look at the equation for the least squares line for  $\text{Log}(\text{HousePrice}(\$1000))$  as a function of  $x = \text{age of house}$ :

$$y' = 2.50 - 0.0087x, \text{ ie}$$

$$\text{Log}(\text{Price}(1000)) = 2.50 - 0.0087 \times \text{Age of House}$$

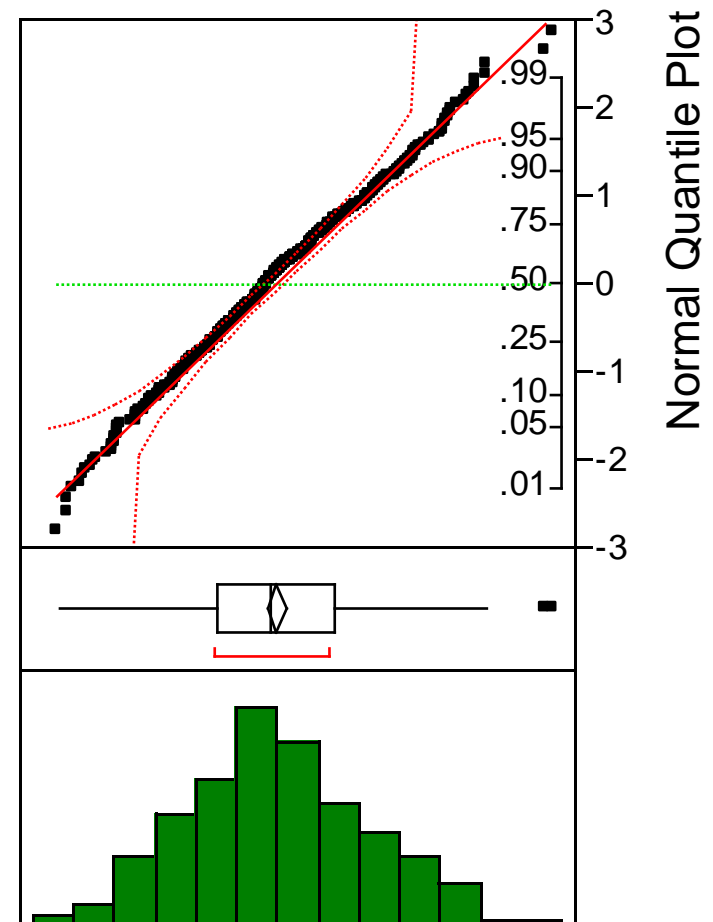
- An annual **decrease** of .0087 in the Log means that the price **decreases** every additional year of age of the house by

$$\frac{y_2}{y_1} - 1 = 100(10^{-0.0087} - 1)\% = -2.0\% .$$

- We have thus *estimated* that HousePrice falls **on average 2%** per year of age of the house.
- We don't know the explanation for this in terms of the housing market.
  - Maybe the quality (and hence price) of houses goes down 2% every year

## 4. Normality

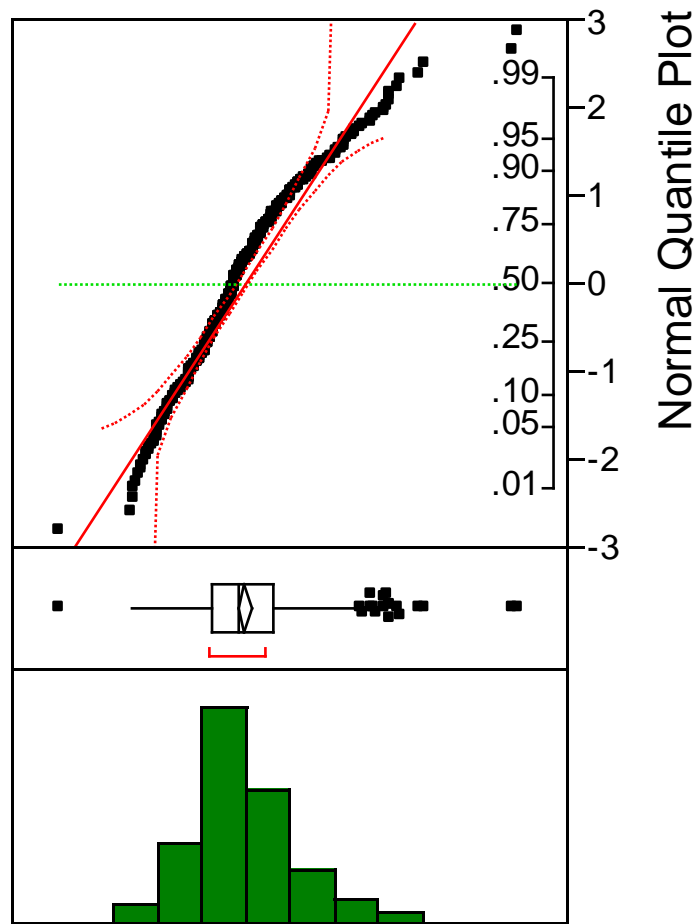
- **IF** you have achieved linearity and homoscedasticity **OR IF** you've done the best you can in that direction **THEN** you should check for normality of residuals
- To do this
  - Save Residuals in JMP.Then examine the Distribution of these residuals. Use the normal quantile plot to check for normality:
- **RESULT: EXCELLENT!**



## What if Normality of Residuals Fails?

- The failure to most worry about is skewness of residuals, or even heavy tails on both sides.
- This type of non-normality may mess up the estimates of the regression line, and will almost surely create somewhat inaccurate confidence intervals.
- As an example of what to look for, here is a picture of skewed residuals.
- This plot is for the residuals of the earlier plot of  
$$Y = \text{HousePrice on } t(x) = \sqrt{x}.$$
- That regression showed a *non-ideal* heteroscedastic pattern that we later fixed, *as we should have*; but here's what the distribution of residuals would have looked like if we had stopped there ----

Residuals corresponding to the LS regression of  $Y$  on  $t(x) = \sqrt{x}$



- Note the skewed to the right pattern evident in both the histogram of residuals and in the normal quantile plot
- Transformations of  $Y$  can sometimes fix skewed residuals

## Transformations of Y to “Fix” Skewness of Residuals Transformations of Y

<i>Group A</i>	<i>Group B</i>
$y'(y) = \sqrt{y}$	$y'(y) = y^2$
$y'(y) = \ln(y)$ or $Log_{10}(y)$	$y'(y) = e^y$
$y'(y) = 1/y$	

- **Group A** transformations may cure skewness to the right.
  - Use  $\sqrt{y}$  for mild skewness and Log for stronger skewness
- **Group B** transformations may cure skewness to the left.
  - Use  $y^2$  for mild skewness and  $e^y$  for stronger skewness
- Use of these transformations may also affect linearity and homoscedasticity, as discussed previously.
- For our data  $Log_{10}(y)$  worked best on all counts!!

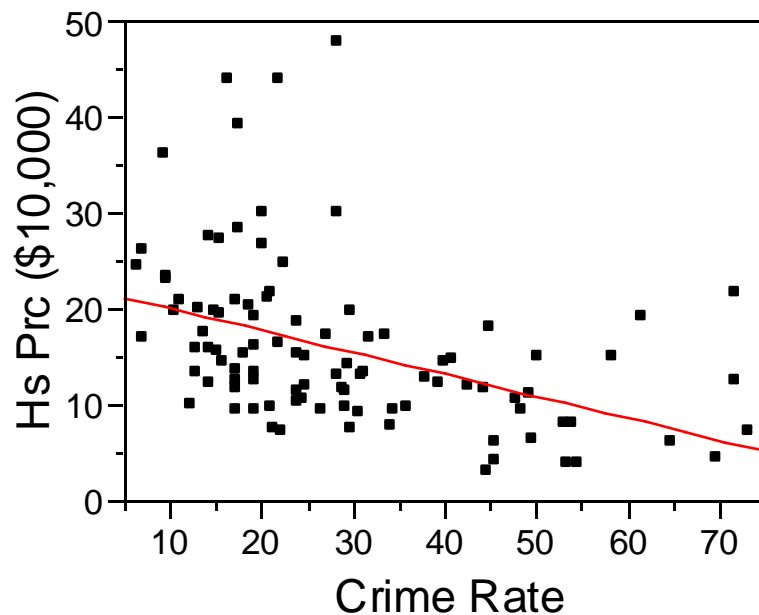
## Influential Point(s)

- We'll show just one example now.
- We'll have opportunities to discuss more examples later.
- Influential points are those whose  $x$ -value lies far from the other  $x$ -values in the data.
- Including the data from such point(s) in the analysis may create a result that heavily depends on just that data point. This may be undesirable.
- Just as with outliers in the vertical (up-down) direction, you may want to think about what these points mean in your data story,
- And you may want to exclude them. If you do so, you may want to tell the story a little differently.

## An Influential Point

Example from the Philadelphia crime data

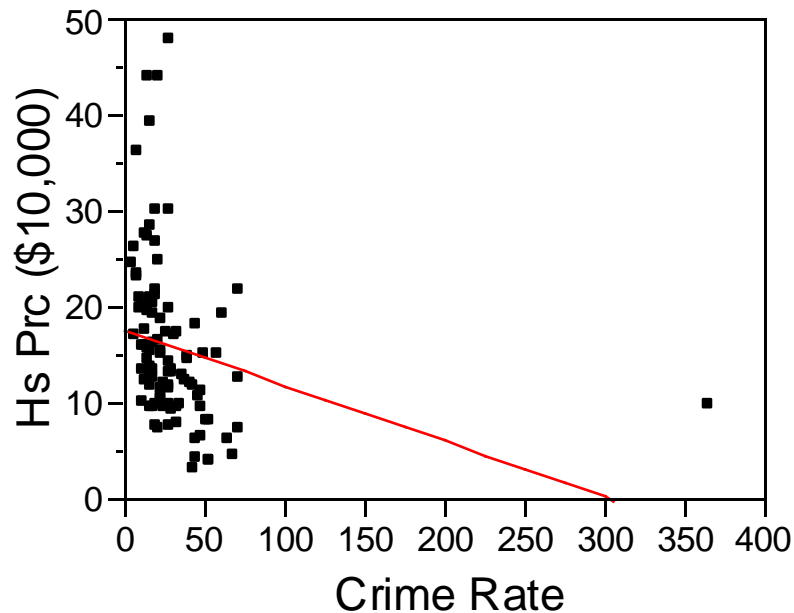
Here's the scatterplot we saw before for this data:



- This scatterplot has some potential outliers (corresponding to “up-scale” suburbs)
- It also has mild non-linearity
- We’ll ignore both of these for now AND look at **all** the data we have

- **All** our data includes Center City Phila and yields ----

## Philadelphia Crime Data INCLUDING Center City Phila.



- For the earlier data (w/o CC Phila) we had  $R^2 = .18$
  - For the data with CC Phila we have  $R^2 = .06$  (much smaller)
  - More important, the slopes are very different:
  - w/o CC  $b_1 = -.23$   
& P-value  $<.0001$
  - with CC  $b_1 = -.06$   
& P-value  $=.013$
- Our proposal is to use the data without CC Phila, and note that our analysis is for Philadelphia communities except for CC