Data Science for Business Analytics (DSBA): Housing Prices and Crime Rates

Subhadip Pal

The Problem

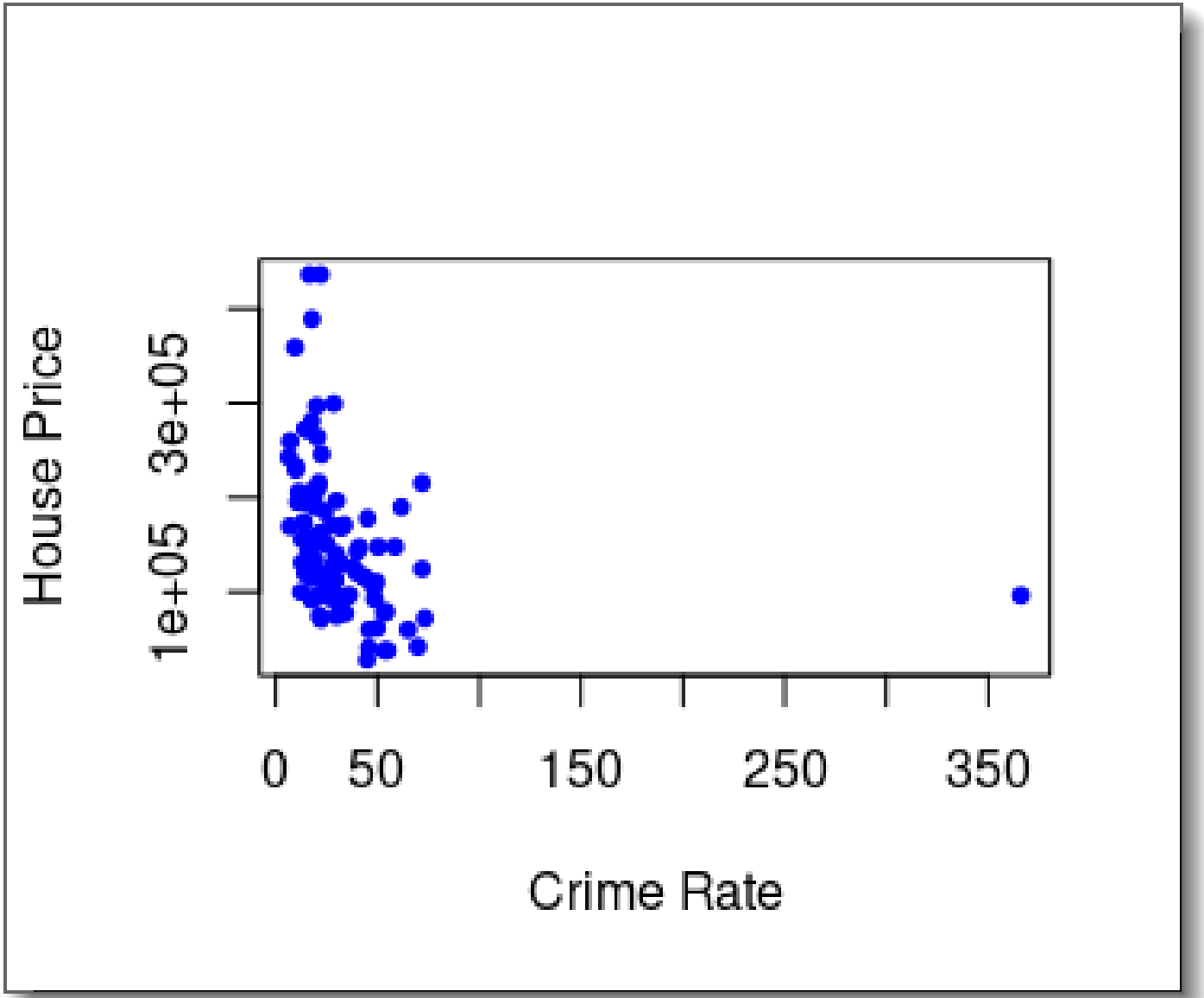
A community in the Philadelphia area is interested in how crime rates affect property values. If low crime rates increase property values, the community might be able to cover the costs of increased police protection by gains in tax revenues from higher property values.

If the community can cut its crime rate from 30 down to 20 per 1000 population, what sort of change might it expect in the property values of a typical house?

The Data

- The city council has data for 110 communities in Pennsylvania near Philadelphia.
- For each community, the data has information on the following:
  HousePrice: average house price during the most recent year
  CrimeRate: Rate of crimes per 1000 population
  MilesPhila: Miles to Philadelphia
  PopChg: Change in population, as a percentage
  Name: Name of the community
- We shall use data for 98 communities, as rest (i.e., 12 communities) have "NA" values for some of the variables
- We shall use the first two variables: *HousePrice* and *CrimeRate*

A First Look at the Data



The Outlier

- This odd community (Center City, Philadelphia) is an outlier on the horizontal axis *CrimeRate*, but is not unusual on the vertical axis (House Price)
- If we fit a least square regression line to this data, the outlier will pull the regression fit toward itself
- As the sum of squared deviations from the regression is minimized in least squares, observations that are far from the fit may have substantial influence on the location of the line
- The size of the impact depends upon the leverage of the outlying value

The Outlier

- Leverage measures how unusal an observation is along the x-axis (here, the *CrimeRate*)
- In simple linear regression, the leverage of the i-th observation is defined as

$$Leverage = h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SS_X},$$

  where SS_X is the sum of squared deviations about the mean of X
- The value of leverage is in the range

$$\frac{1}{n} \leq h_i \leq 1$$

- Here, Center City, Philadelphia is a highly leveraged observation (leverage = 0.82)

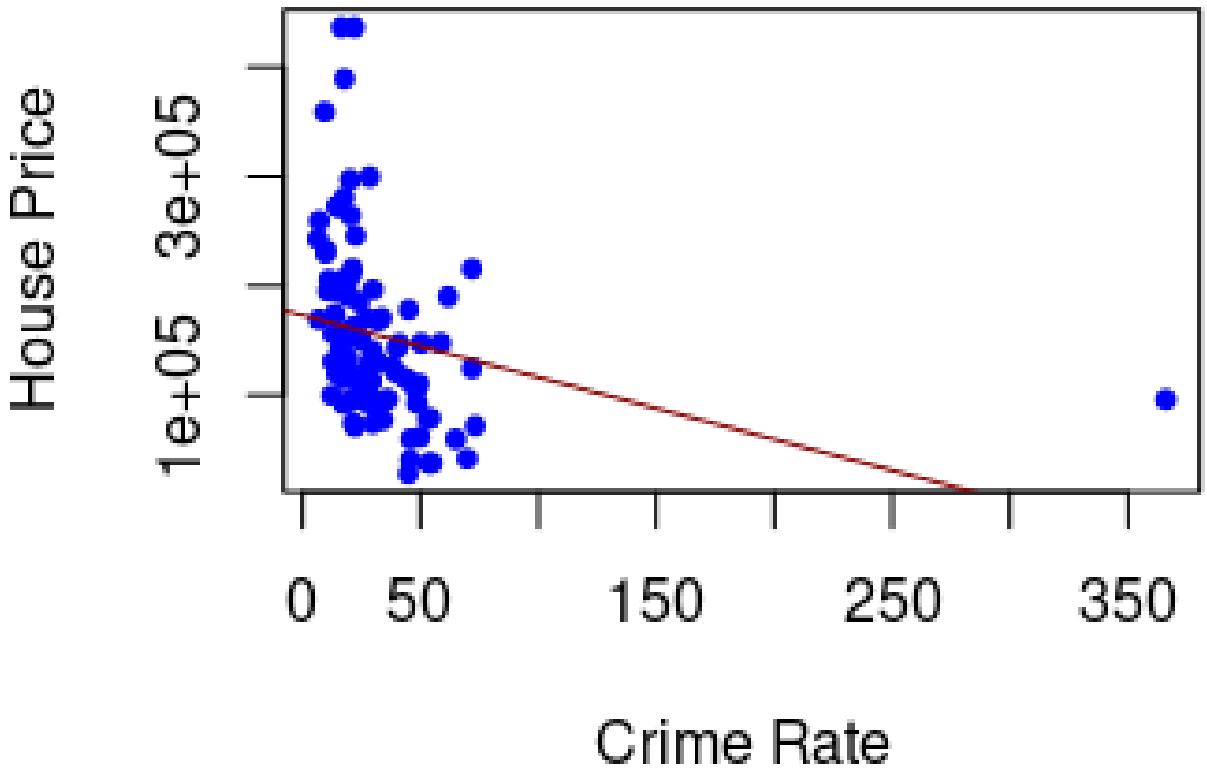
Fitting a Simple Regression Model

- First, let us fit a simple regression model to the data
- *HousePrice* is the response, and *CrimeRate* is the predictor

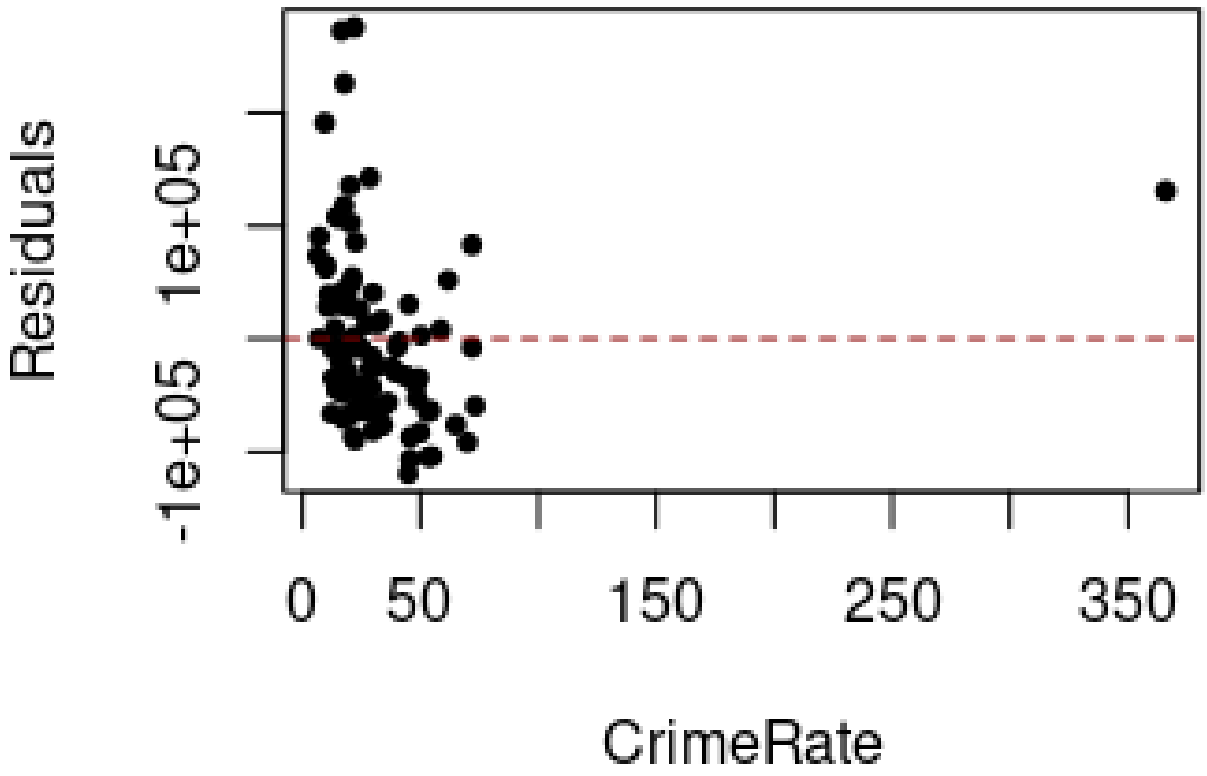
Table: Simple Regression Model for the Data

	Estimate	Std. Error	*t*-value	*p*-value
Intercept	173116.4	10486.5	16.509	$<$ 2e-16
CrimeRate	-567.7	210.9	-2.692	0.00837

The Fitted Line and the Data

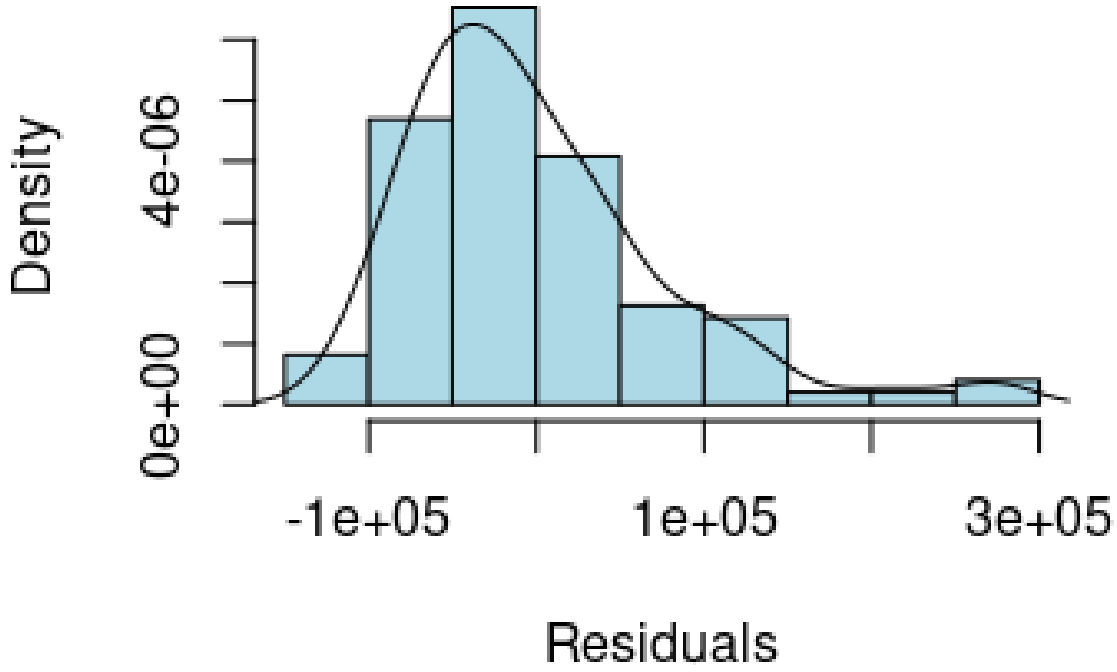


Residual Plot



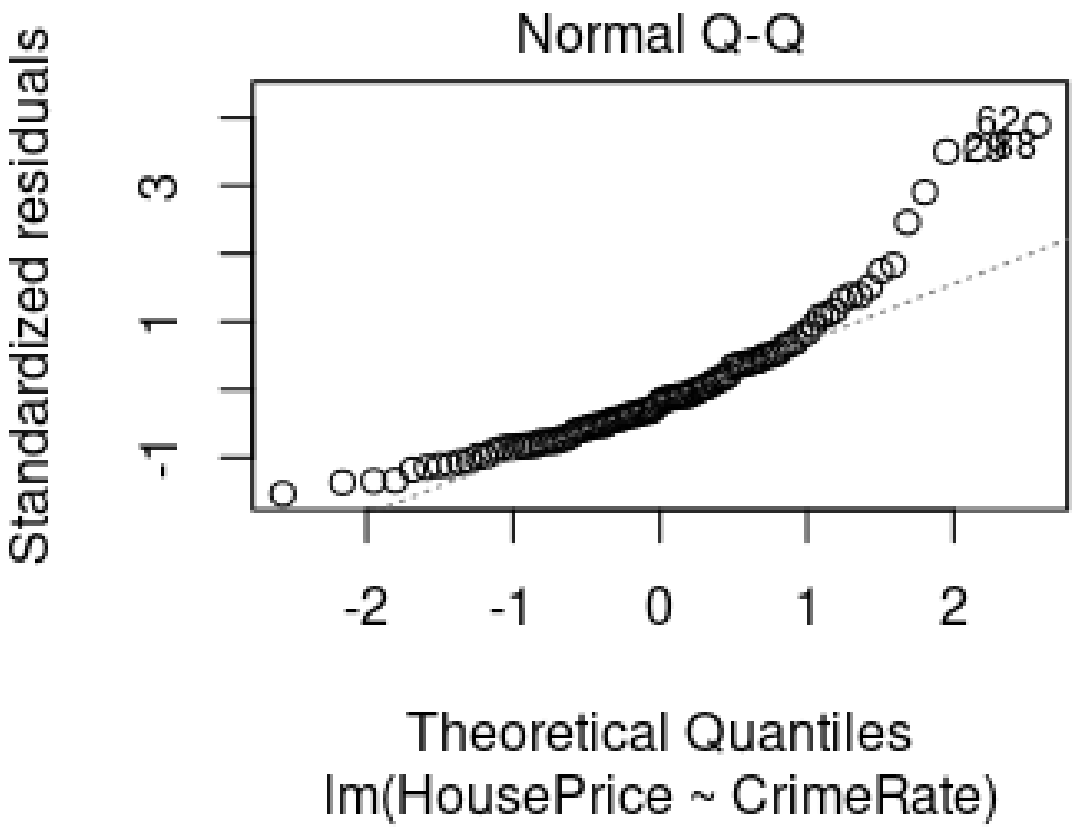
Distribution of Residuals

- The overall distribution of the residuals is skewed to the right



The QQ Plot

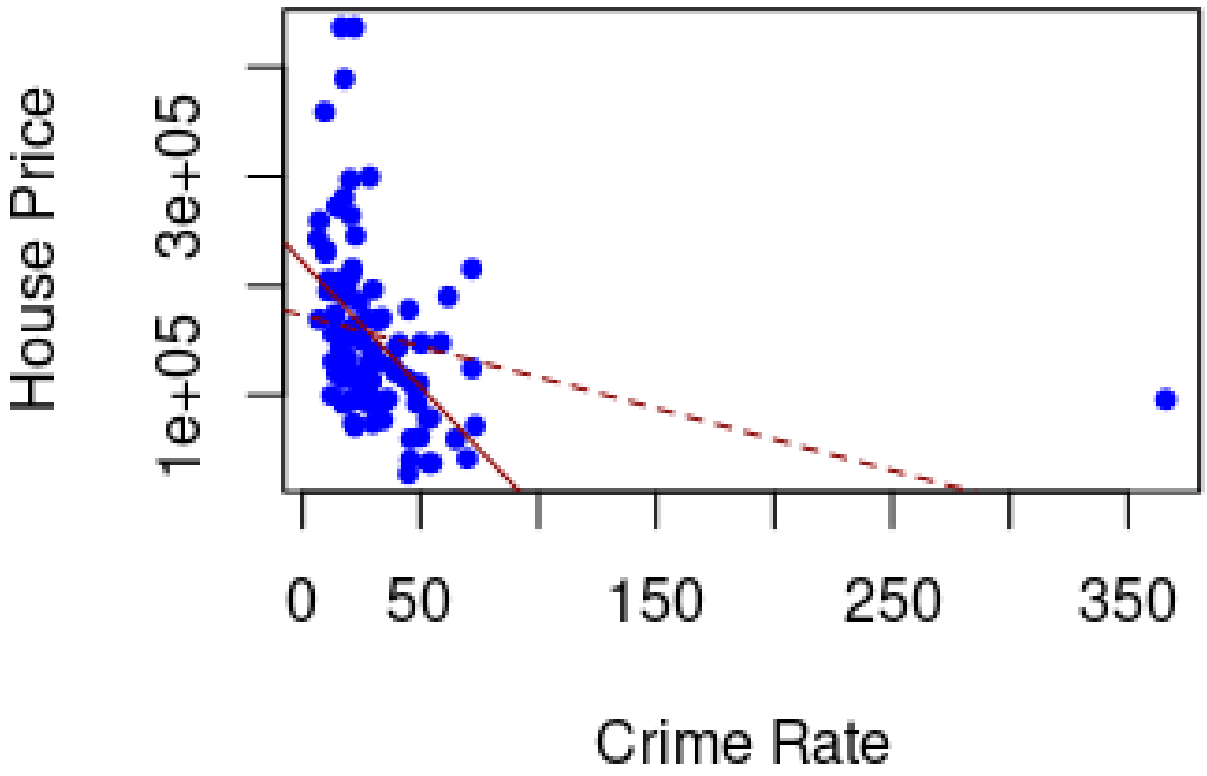
- The Quantile-Quantile (QQ) plot of the residuals shows that there is a problem with normality assumption, although the plot is dominated by the outlier



Handling the Outlier

- It is clear from the above analyses that the outlier is affecting the model quite strongly
- We need to handle the outlier before further analyses
- Let us repeat the analyses deleting the outlier

A Model Without The Outlier



A Model Without The Outlier

- Delete the outlier, and fit the simple regression model as before

Table: Simple Regression Model for the Data

	Estimate	Std. Error	*t*-value	*p*-value
Intercept	221775.6	15059.6	14.727	$<$ 2e-16
CrimeRate	-2281.6	450.6	-5.063	2.02e-06

Comparison: Two Models

- The regression model with the outlier in the data is

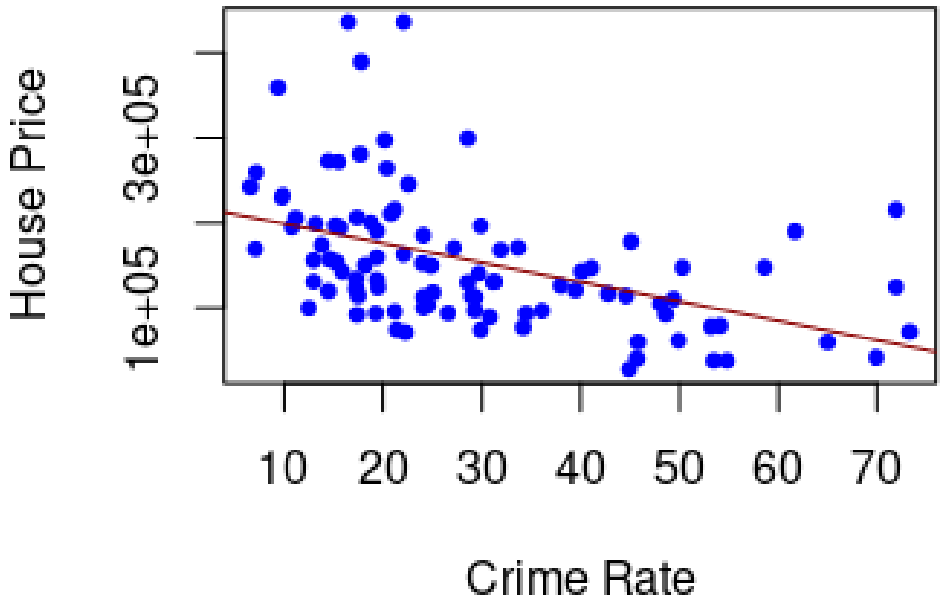
$$HousePrice = 173116.4 - 567.7 \times CrimeRate$$

- The regression model without the outlier in the data is

$$HousePrice = 221775.6 - 2281.6 \times CrimeRate$$

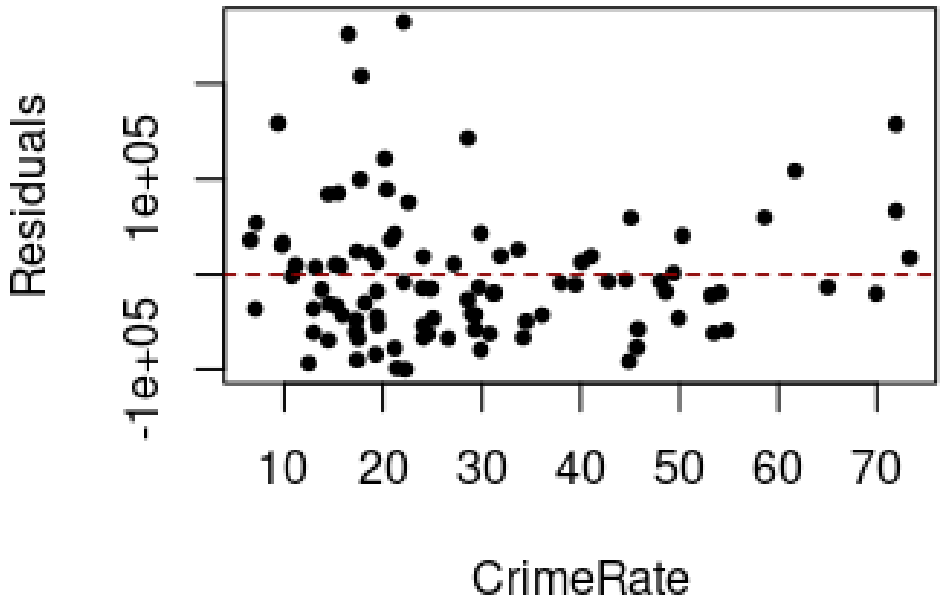
- Without the outlier, the slope of the regression line is almost four times the slope of the previous line
- The change in the regression line is dramatic, and so is the change in the estimated effect of *Crime Rate* on *House Price*

The New Fit



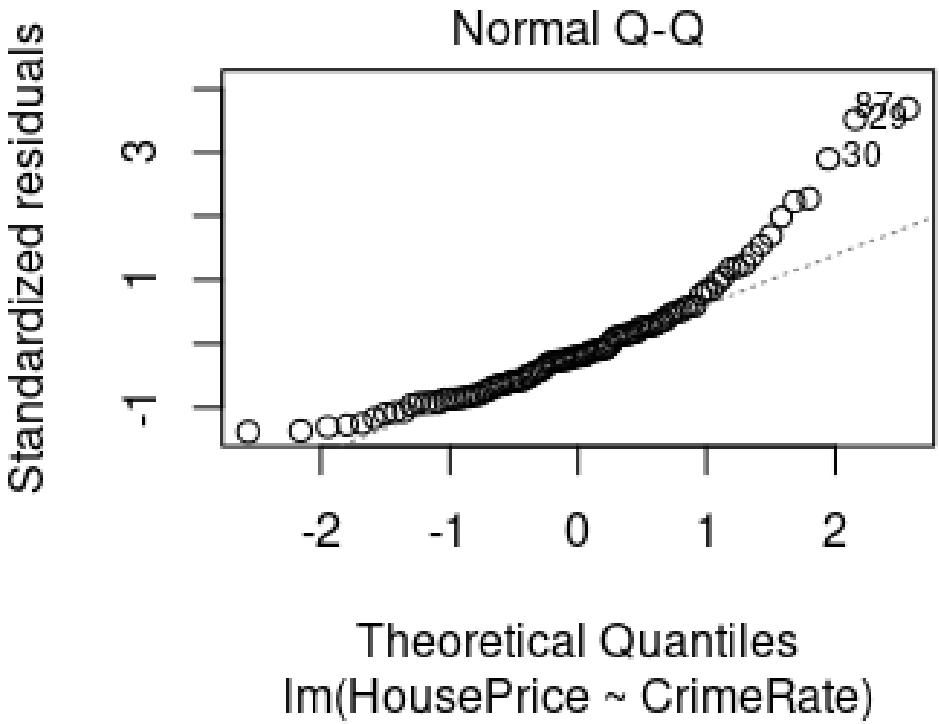
- While the fit has definitely improved, but the plotted points seem to show a nonlinear pattern

New Residual Plot



- Residuals also indicate nonlinearity; implies some transformation of the predictor will be useful

New QQ Plot



- The QQ plot clearly shows positive skewness of the residuals (i.e., many small negative residuals, and few large positive residuals)

Impact of Crime Rate

- The regression model with Center City in the data is

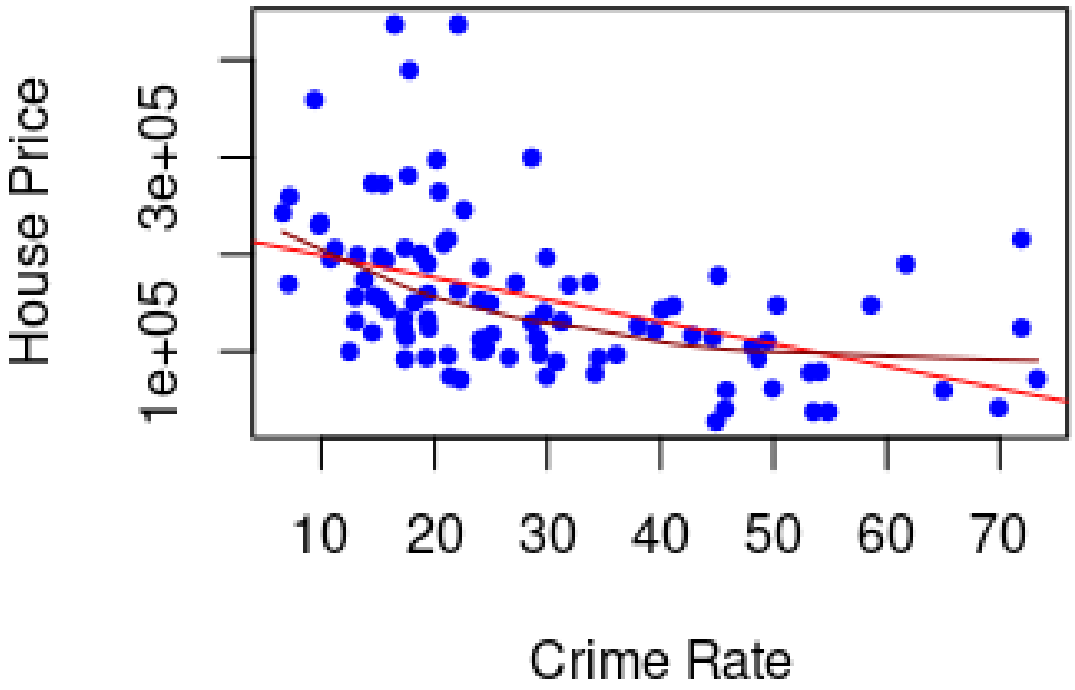
$$HousePrice = 173116.4 - 567.7 \times CrimeRate$$

- The regression model without Center City in the data is

$$HousePrice = 221775.6 - 2281.6 \times CrimeRate$$

- Consider a 10-point drop in *CrimeRate*
- With Center City, the increase in *HousePrice* is $10 \times \$567.7 \approx \$5700$ per house
- Without Center City, the increase in *HousePrice* is $10 \times \$2281.6 \approx \$23000$ per house
- Caution: The data for house prices are from sales of current year only - may not be representative of the actual sales

Scatterplot Smoothing



- The nonlinear relationship is clear!

Fitting a Transformed Model

- We fit the reciprocal model to this data, that is,

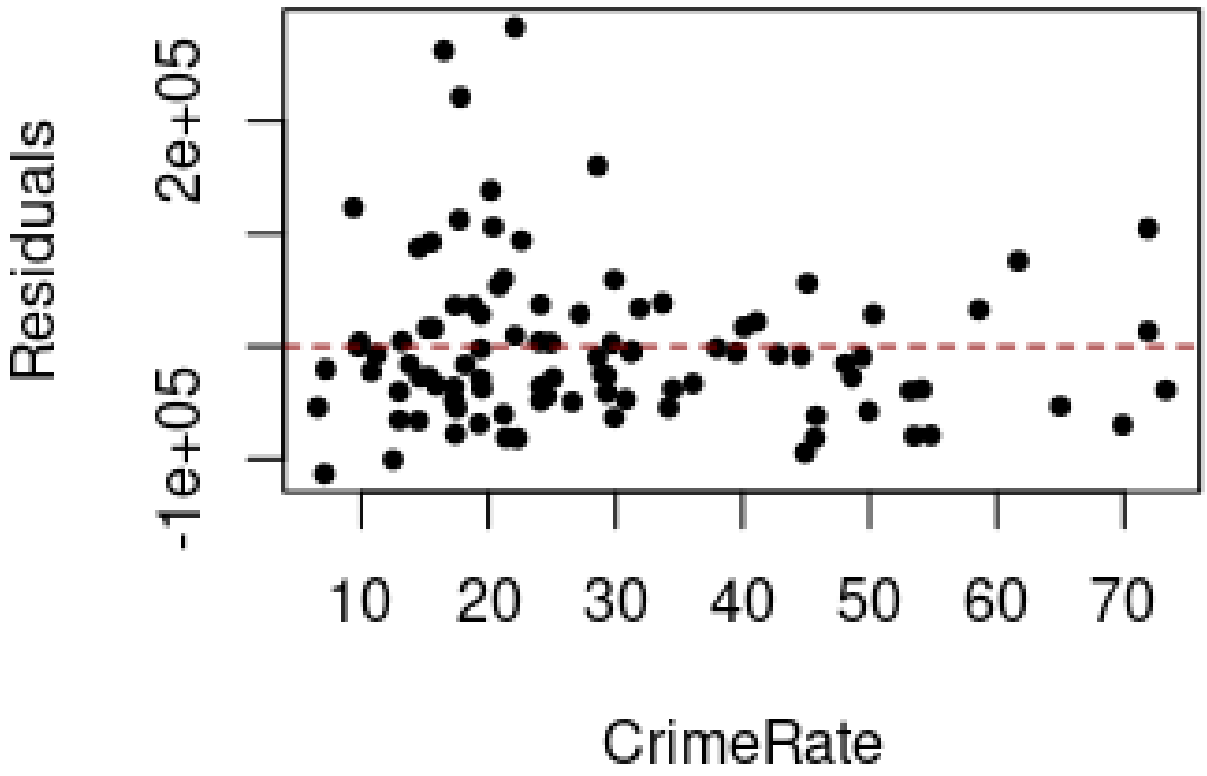
$$HousePrice = \alpha + \beta \times (1/CrimeRate)$$

- With the estimated coefficients, the model is

$$HousePrice = 92256 + 1350722 \times (1/CrimeRate)$$

- To see if the new transformed model is better on this data, look at the residuals - a plot of the residuals against *CrimeRate*, and the QQ plot of the residuals

Residual Plot for The Transformed Model



Impact of Crime Rate: The Transformed Model

- Change in fit(*CrimeRate* 20 to 10) $= \frac{13350722}{10} - \frac{13350722}{20} = \$67536$
- Change in fit(*CrimeRate* 30 to 20) $= \frac{13350722}{20} - \frac{13350722}{30} = \$22512$
- The fitted model flatens out as *CrimeRate* increases
- That is, increase in *CrimeRate* has a big effect on house prices for smaller crime rates; the effect is smaller when the crime rates are larger

A Word of Caution

- Do crime rates affect house prices, or is it the other way round?
- Maybe some third factor, like education? (lurking variable)
- Regression, like correlation, is based on association
- In general, it cannot deduce cause and effect relationship (unless it is based on an experimental design)

Thank You

- Thank You