STAT 380:
Data: Housing Prices and Crime Rates

(Check the R code from the file : HousingPrice.Rmd)

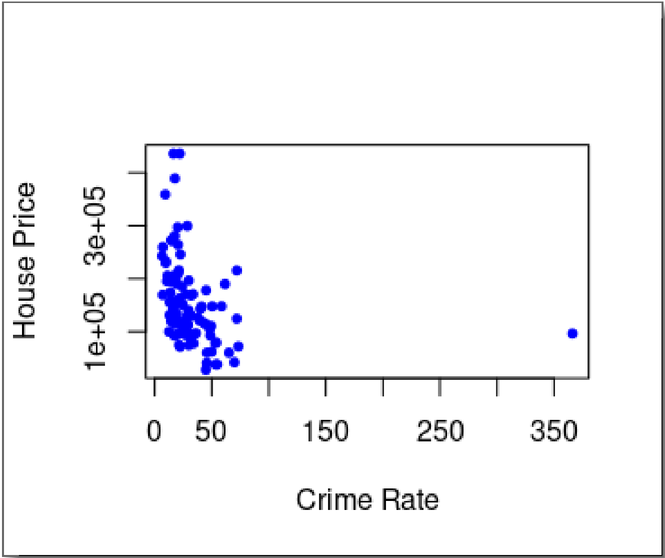
The Problem

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*


House Price
1e+05
3e+05
0
50
150
250
350
Crime Rate

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
- The size of the impact depends upon the influential identified by the Cook's Distance.

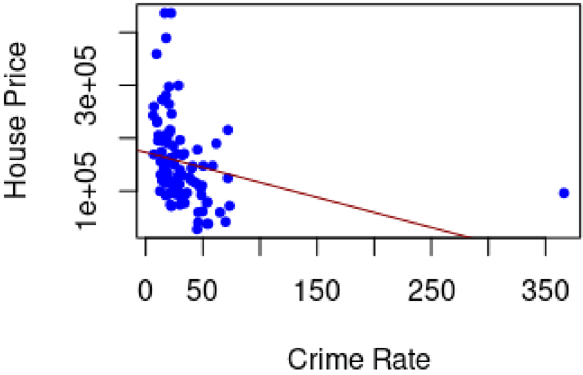
Fitting a Simple Regression Model

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

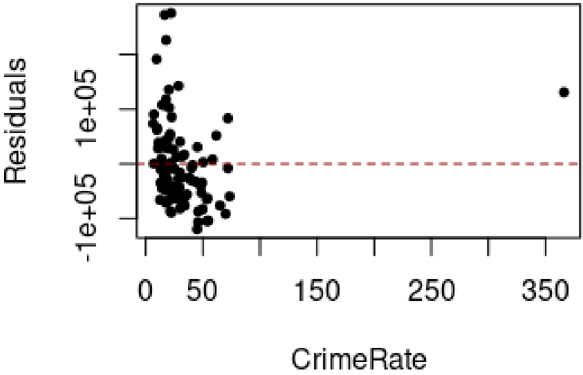
Table: Simple Regression Model for the Data

	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

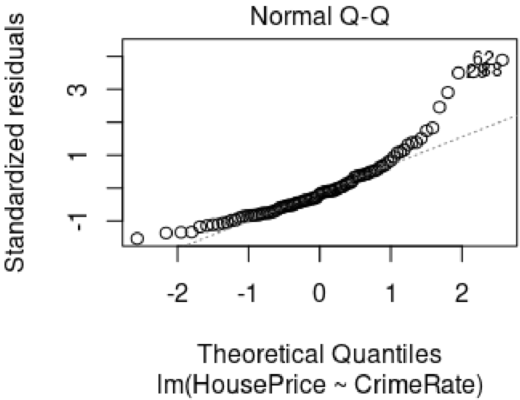


Residual Plot



Test of Normality: The QQ Plot

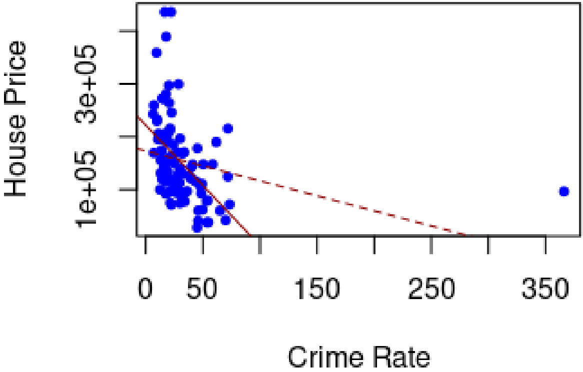
- The Quantile-Quantile (QQ) plot of the residuals shows that there is a problem with normality assumption, although the plot is dominated by the outlier



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/Influential Points



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

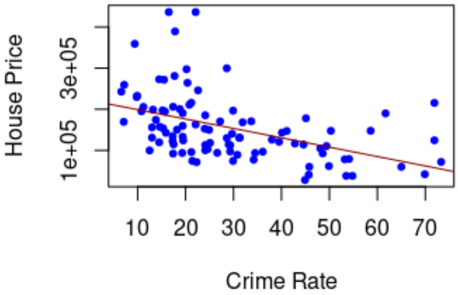
- The regression model with the outlier in the data is

$$\textit{HousePrice} = 173116.4 - 567.7 \times \textit{CrimeRate}$$

- The regression model without the outlier in the data is

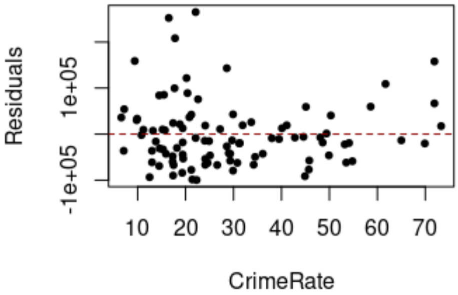
$$\textit{HousePrice} = 221775.6 - 2281.6 \times \textit{CrimeRate}$$

- 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*



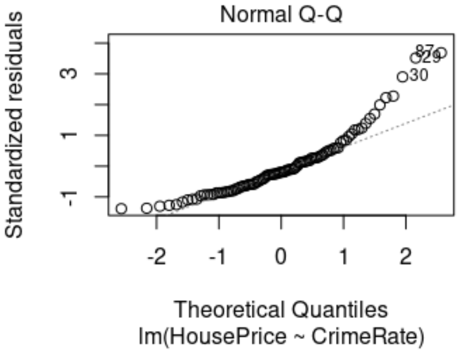
- While the fit has definitely improved, but the plotted points seem to show a nonlinear pattern

New Residual Plot



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

QQ Plot



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

Shapiro-Wilk normality test for testing normality of the residuals:
The test Statistic: $W = 0.88802$,
The corresponding p-value < 0.00001

Fitting a Transformed Model

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

$$\log(\mathit{HousePrice}) = \beta_0 + \beta_1 \times (\mathit{CrimeRate}) + \epsilon$$

- With the estimated coefficients, the model is

$$\log(\mathit{HousePrice}) = 12.325760 - 0.01730 \times (\mathit{CrimeRate})$$

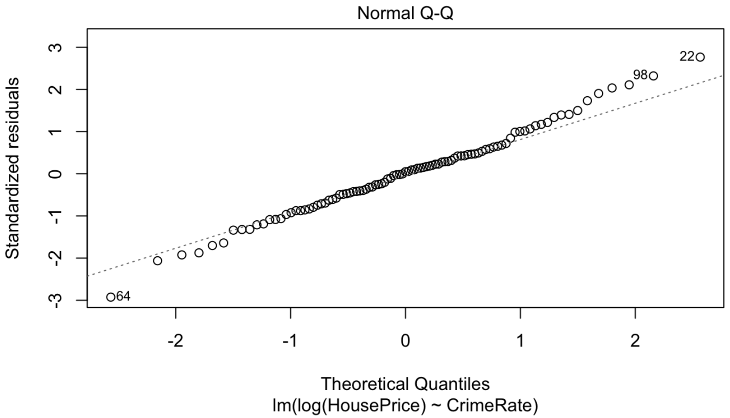
$$\mathit{Estimated} : \mathit{HousePrice} = e^{12.325760 - 0.01730 \times (\mathit{CrimeRate})}$$

- 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

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)

QQ Plot From Model 3



Shapiro-Wilk normality test for testing normality of the residuals:
The test Statistic: $W = 0.99375$,
The corresponding p-value $= 0.9359$

- Thank You