Analysis of Economics Data
Chapter 5: Bivariate Data Summary

© A. Colin Cameron
Univ. of Calif. Davis

November 2022

CHAPTER 5: Bivariate Data Summary

- Summarize the relationship between two variables:
  - e.g. earnings and education
  - e.g. house price and house size
  - notation is that variable y is a function of variable x.
- How do we measure the association?
  - **correlation coefficient** $-1 \leq r \leq 1$.
- How do we summarize the relationship?
  - **linear regression** $\widehat{y} = b_1 + b_2 x$.
- How do we summarize the strength of this relationship?
  - **R-squared** $0 \leq R^2 \leq 1$.
  - standard error of the regression s_e.
- This chapter provides details on these measures.

Outline

1. Example: House Price and Size
2. Two-way Tabulation
3. Two-way Scatter Plot
4. Correlation
5. Regression Line
6. Measures of Model Fit
7. Computer Output following OLS Regression
8. Prediction and Outlying Observations
9. Regression and Correlation
10. Causation
11. Computations for Correlation and Regression
12. Nonparametric Regression

Dataset: HOUSE.

5.1 Example: House Price and Size

- House price and size for sample of 29 houses
  - control for location by consider homogeneous housing market
  - central Davis in 1999.
  - eyeballing data it seems higher price if larger size

Sale Price	Sq. Feet	Sale Price	Sq. Feet	Sale Price	Sq. Feet
375,000	3,300	255,000	1,500	235,000	1,700
340,000	2,400	253,000	2,100	233,000	1,700
310,000	2,300	249,000	1,900	230,000	2,100
279,900	2,000	245,000	1,400	229,000	1,700
278,500	2,600	244,000	2,000	224,500	2,100
273,000	1,900	241,000	1,600	220,000	1,600
272,000	1,800	239,500	1,600	213,000	1,800
270,000	2,000	238,000	1,900	212,000	1,600
270,000	1,800	236,500	1,600	204,000	1,400
258,500	1,600	235,000	1,600		

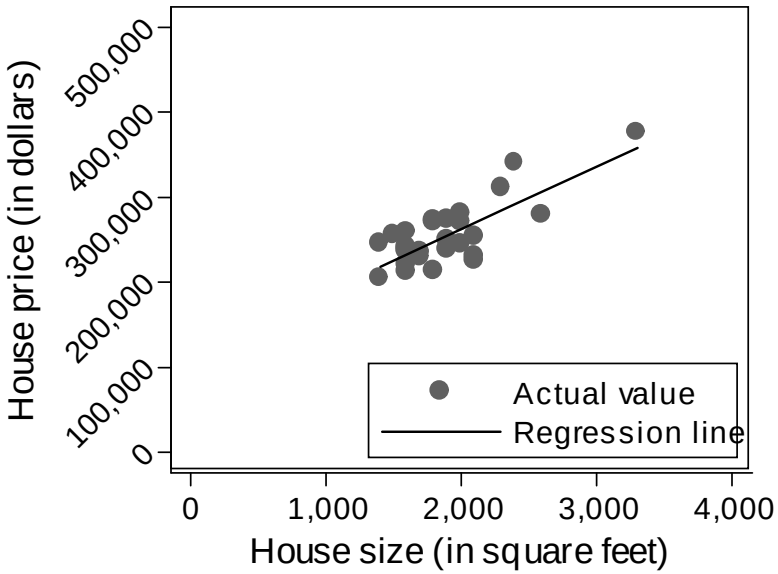
Summary Statistics

- House sale price ranges from \$204,000 to \$375,000
  - mean \$253,910 and standard deviation \$37,391.
- House size ranges from 1,400 to 3,300 square feet
  - mean 1,883 square feet and standard deviation 398 square feet.

Statistic	Sale Price	Square Feet
Mean	253,910	1,883
Standard deviation	37,391	398
Standard error	6,943	74
Maximum	375,000	3,300
Median (50th percentile)	244,000	1,800
Minimum	204,000	1,400
Skewness	1.56	1.73
Kurtosis	5.61	6.74

Key methods for measuring relationship (this chapter)

- The correlation between house price against house size is 0.786.
- A scatterplot of house price against house size yields



- Regression of house price against house size yields

$$\widehat{Price} = 115,017 + 73.77 \times Size. \qquad R^2 = 0.6175$$

- An extra square foot of house is associated with a \$73.77 increase in house price.

5.2 Two-way Tabulation

- A **two-way tabulation** or **cross tabulation** of variables x and y lists the number (or fraction) of observations equal to each of the distinct values taken by the pair (x, y).
- Useful if the variables x and y take relatively few values
  - categorical data with few categories
  - discrete numerical taking a few values
  - for continuous numerical convert to a few ranges.
- House price and size data create
  - *pricerange:* low (*price*$<$\$249,000) or high (*price*$\geq$\$250,000).
  - *sizerange:* small (*size*$<$1,800), medium (1,800$\leq$*size*$<$2,400) or large (*size*$\geq$2,400).

Two-way Tabulation (continued)

- Main entry: # observations with a given price-size combination
  - e.g. there were 11 houses of low price and small size.

	Size range			
Price range	Small	Medium	Large	Total
Low	11	6	0	17
High	2	7	3	12
Total	13	13	3	29

- Table also includes row sums and column sums
  - e.g. total in row for low price range is $11 + 6 + 0 = 17$ observations.
- Table includes a second optional entry, a **row percentage**
  - for each value of *pricerange* gives % of obs in each of the size ranges
  - e.g. low-priced: small $= 11$ out of $17 = 100 \times 11/17 = 64.71\%$.
- Can also include similarly constructed **column percentages**.

Two-way Tabulation (continued)

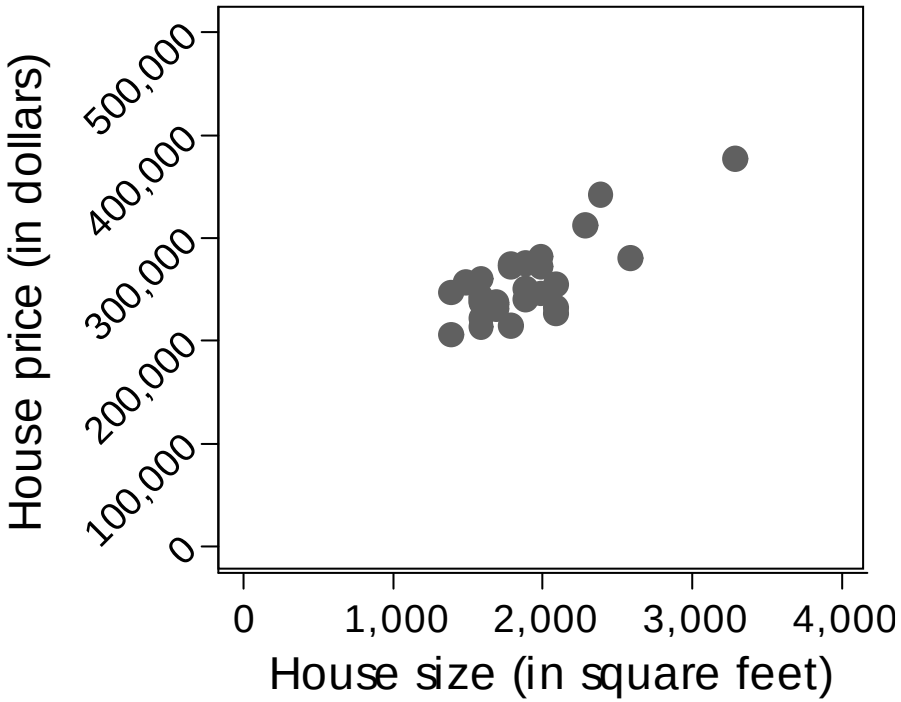
- A two-way tabulation can also include **expected frequencies**, assuming that the two variables are statistically independent.
  - Expected frequency = (row total x column total) / # obs.
  - e.g. low-price small-size cell expect $17 \times 13/29 = 7.62$.

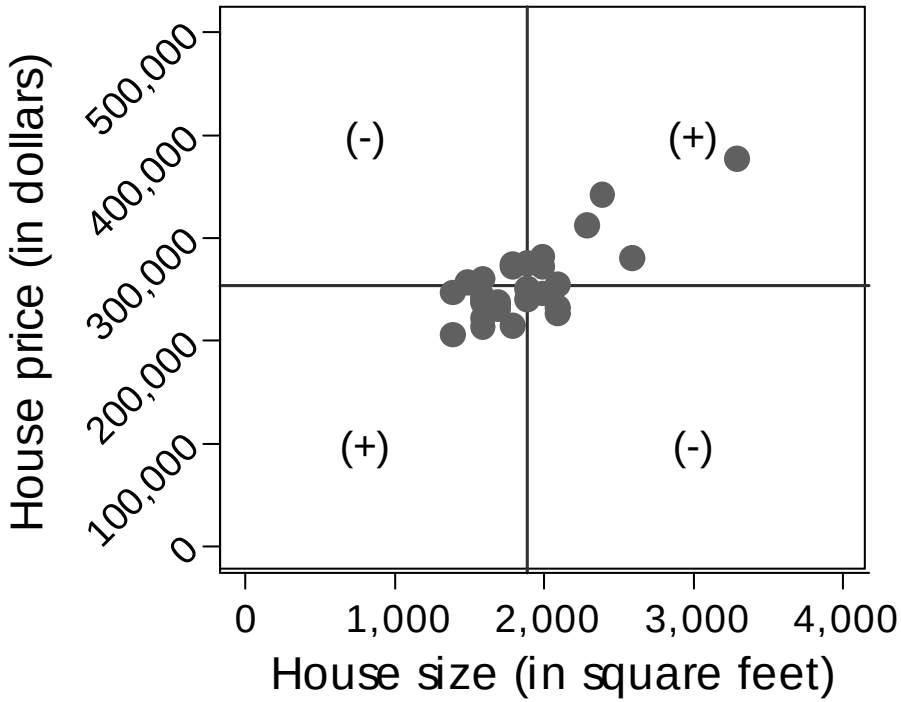
	Size range			
Price range	Small	Medium	Large	Total
Low	11	6	0	17
	7.62	7.62	1.76	17.00
High	2	7	3	12
	5.38	5.38	1.24	12.00
Total	13	13	3	29
	44.83	44.83	10.34	29.00

- Table presents both observed and expected frequencies.
  - e.g. More low-price houses are small then would be expected if price and size were independent (11 versus 7.62)
  - Difference is basis for Pearson's chi-squared goodness-of-fit test of statistical independence of two categorical variables.

5.3 Two-way Scatterplot

- Standard visual method is a two-way scatter plot
  - first panel shows house price increases with house size.





5.4 Sample Correlation

- **Correlation coefficient** is a standard way to measure association between x and y
- The **sample correlation coefficient** is a unit-free measure ranging from -1 to 1 with

$r_{xy} = 1$	perfect positive correlation
$0 < r_{xy} < 1$	positive correlation
$r_{xy} = 0$	no correlation
$-1 < r_{xy} < 0$	negative correlation
$r_{xy} = -1$	prefect negative correlation

- For the house price and size data: $r_{xy} = 0.786$.

Sample Covariance

- Recall sample variance $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
- The **sample covariance** between x and y is similarly defined:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

- Suppose on average y increases as x increases.
  - then $(x_i - \bar{x})(y_i - \bar{y}) > 0$ most of the time
    - ★ since $(y_i - \bar{y}) > 0$ usually if $(x_i - \bar{x}) > 0$, so $(+)\times(+) = (+)$
    - ★ and $(y_i - \bar{y}) < 0$ usually if $(x_i - \bar{x}) < 0$, so $(-)\times(-) = (+)$
  - It follows that $s_{xy} > 0$.
- Example is the second panel on the earlier slide
  - Most observations are in the quadrants where $(x_i - \bar{x})(y_i - \bar{y}) > 0$
  - so positively correlated (in fact $s_{xy} = 11{,}701{,}613.3$!).
- Similarly $s_{xy} < 0$ if y decreases as x increases.
- Thus the sign of the covariance is easily interpreted
  - $s_{xy} > 0$ if positive association
  - $s_{xy} < 0$ if negative association.

Sample Correlation

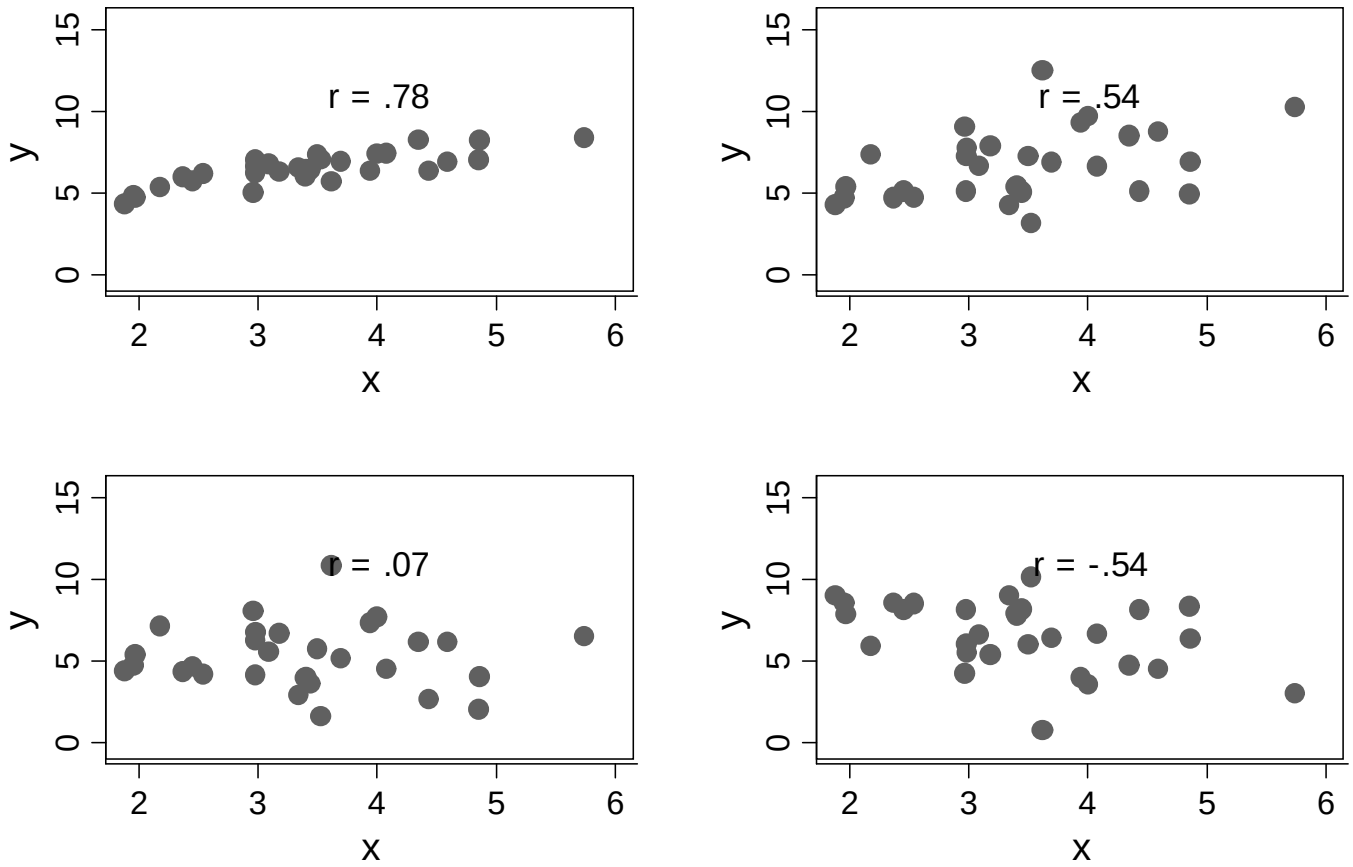
- The **sample correlation coefficient** is defined by

$$\begin{aligned} r_{xy} &= \frac{\text{Covariance of } x \text{ and } y}{(\text{Standard deviation of } x) \times (\text{Standard deviation of } y)} \\ &= \frac{s_{xy}}{s_x s_y} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \end{aligned}$$

- The correlation coefficient is the covariance between the standardized versions of x and y
  - r_{xy} equals the covariance of $(x - \bar{x})/s_x$ and $(y - \bar{y})/s_y$.

Four Examples of Strength of Correlation

- (1) strong positive correlation; (2) moderate positive correlation; (3) almost zero correlation, and (4) moderate negative correlation.
- Though no clear cutoffs for "weak", "moderate", "strong" correlation.



Autocorrelations for Time Series Data

- For time series data the **autocorrelation** at lag j is the correlation between current data and the data lagged j periods.
  - e.g. correlation between y_t and y_{t-j}.

5.5 Regression Line

- This is the key method in the analysis of economics data.
- The **regression line** from regression of y on x is denoted

$$\widehat{y} = b_1 + b_2 x,$$

where

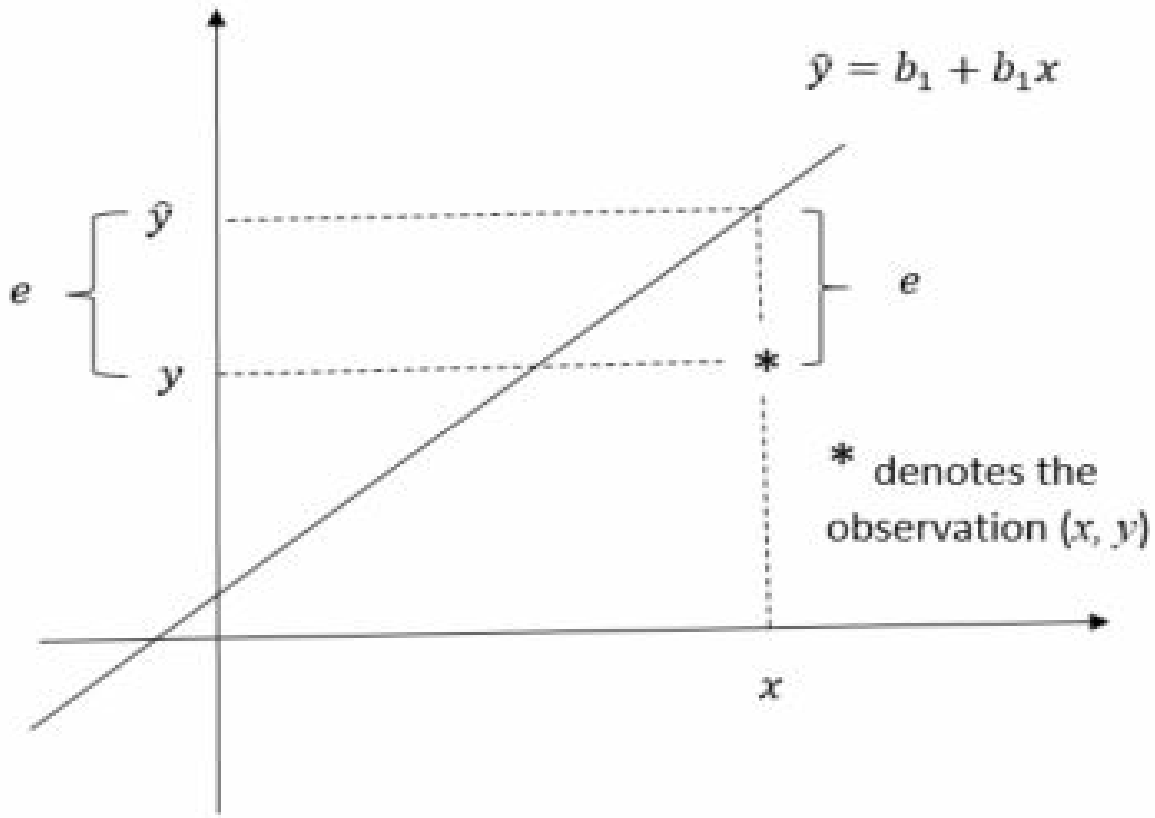
- y is called the **dependent variable**
- $\widehat{y}$ is the **predicted value** or **fitted value** of the dependent variable
- x is the **independent variable** or **explanatory variable** or **regressor variable** or **covariate**
- b_1 is the estimated y-axis **intercept**
- b_2 is the estimated **slope** coefficient.

The Residual

- **Residual** e is the difference between actual value of y and predicted value $\widehat{y}$

$$e = y - \widehat{y}.$$

  - also denoted $\widehat{u} = y - \widehat{y}$.



Least Squares Regression

- For first observation the residual is $e_1 = y_1 - \widehat{y}_1$, for second observation the residual is $e_2 = y_2 - \widehat{y}_2$, and so on.
- For i^{th} observation

$$\begin{aligned} e_i &= y_i - \widehat{y}_i \\ &= y_i - b_1 - b_2 x_i. \end{aligned}$$

- **Least squares method** chooses intercept b_1 and slope b_2 of the line to **make as small as possible the sum of the squared residuals,** $e_1^2 + e_2^2 + \cdots + e_n^2$.
- Thus b_1 and b_2 minimize

$$\begin{aligned} \sum_{i=1}^n e_i^2 &= \sum_{i=1}^n (y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - b_1 - b_2 x_i)^2. \end{aligned}$$

- This is a calculus problem
  - Differentiate with respect to b_1 and b_2
  - Set the two derivatives equal to zero
  - Solve two equations in two unknowns for b_1 and b_2
  - Algebra is skipped.
- The resulting formula for the least squares **slope coefficient** is

$$b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- The least squares **intercept** is

$$b_1 = \bar{y} - b_2\bar{x}.$$

Interpretation of the Slope Coefficient

- The slope coefficient b_2 gives the slope:

$$\frac{\Delta \widehat{y}}{\Delta x} = b_2.$$

- Reason: If regressors changes by Δx from x to $(x + \Delta x)$ then the fitted value $\widehat{y}$ changes from $b_1 + b_2 x$ to $b_1 + b_2(x + \Delta x) = b_1 + b_2 x + b_2 x \Delta x$, a change of $b_2 \Delta x$.
  - It follows that $\Delta \widehat{y} = b_2 \Delta x$.
- The slope coefficient b_2 is therefore easily interpreted as the change in the predicted value of y when x increases by one unit.
- The same result can be obtained using calculus methods
  - since $\widehat{y} = b_1 + b_2 x$ has derivative $d\widehat{y}/dx = b_2$.

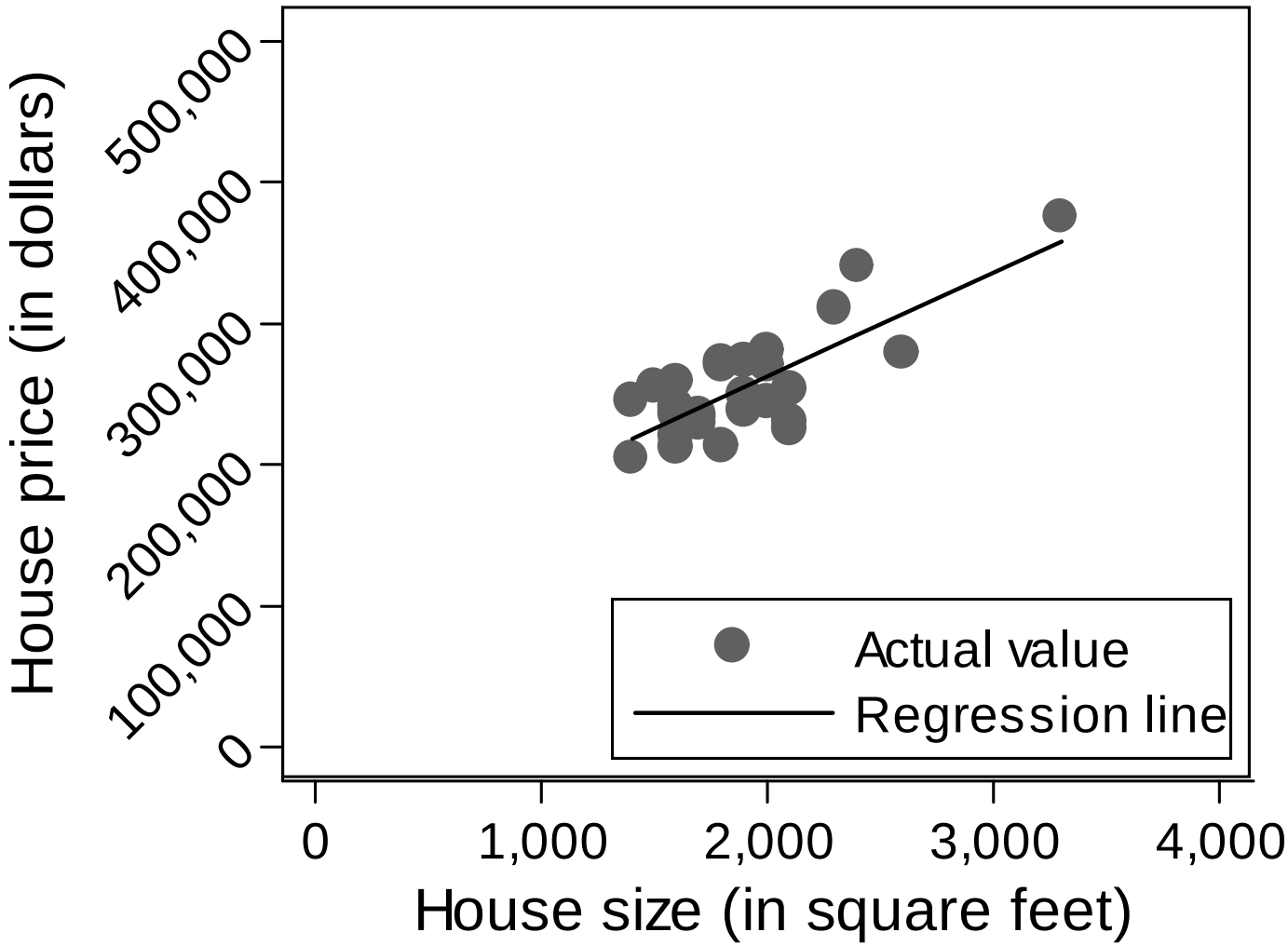
Example: House Price

- Fitted regression

$$\widehat{Price} = 115,017 + 73.77 \times Size.$$

- The slope coefficient equals 73.77
  - one more square foot in size is associated with a \$73.77 increase in the house price
  - equivalently an additional small room of size ten feet by ten feet, or 100 square feet, is associated with a $100 \times \$73.77 = \$7,377$ increase in house price.

- Scatterplot and least squares regression line.



Intercept-only Regression yields Sample Mean

- OLS regression of y on just an intercept
  - minimize $\sum_{i=1}^{n}(y_i - b_1)^2$ yields $b_1 = \bar{y}$.
- So regression of y on only an intercept yields the sample mean $\bar{y}$
- OLS regression is a natural extension of univariate statistics based on the sample mean
- And univariate statistics based on the sample mean is just a special case of OLS regression.

5.6 Measures of Model Fit

- Two standard measures.
- The standard error of the regression measures the standard deviation of the residuals.
- R-squared (R^2) measures the fraction of the variation of y (around the sample mean $\bar{y}$) that is explained by the regressors.
- Provided the regression includes an intercept
  - $0 \leq R^2 \leq 1$
  - $R^2 = 0 \Rightarrow$ no relationship between y and x as $\widehat{y}_i = \bar{y}$ for all i.
  - $R^2 = 1 \Rightarrow$ regression line perfectly fits y as $\widehat{y}_i = y_i$ for all i.
  - $R^2 = r_{xy}^2 \Rightarrow R^2$ equals the squared correlation coefficient
  - $R^2 =$ the squared correlation between y and x (i.e. $R^2 = r_{xy}^2$)
  - $R^2 =$ the squared correlation between y and fitted values $\widehat{y}$.

Standard Error of the Regression

- The **standard error of the regression** is

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

  - also called the root mean squared error of the residual.
- This measures the closeness of the fitted values $\widehat{y}_i$ to the actual values y_i
  - it is essentially the average of the squared residuals
    - ★ except that division is by $n-2$ rather than n.
- Lower values of s_e means fitted values are closer to actual values
  - but s_e is not scale free.

Definition of R-squared

- R^2 measures the fraction of the variation of y (around the sample mean $\bar{y}$) that is explained by the regressors.
- **Total sum of squares:** measures variability in y around the sample mean $\bar{y}$

$$Total\ SS = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

- **Explained sum of squares** measures variability in fitted value $\widehat{y}$ around $\bar{y}$

$$Explained\ SS = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2$$

  - also called **regression sum of squares** or **model sum of squares**.

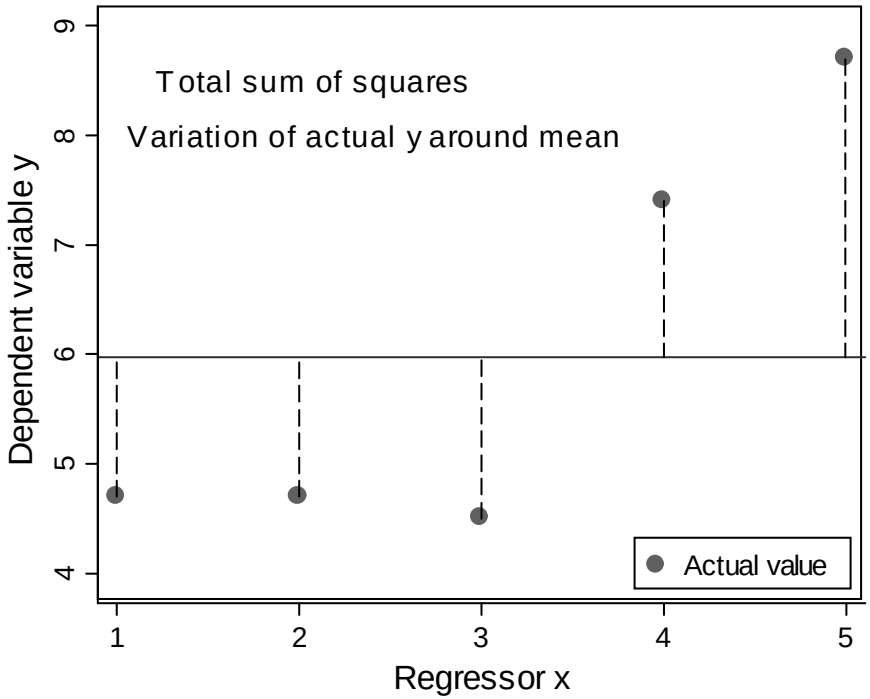
- **R-squared** equals explained sum of squares as a fraction of the total sum of squares

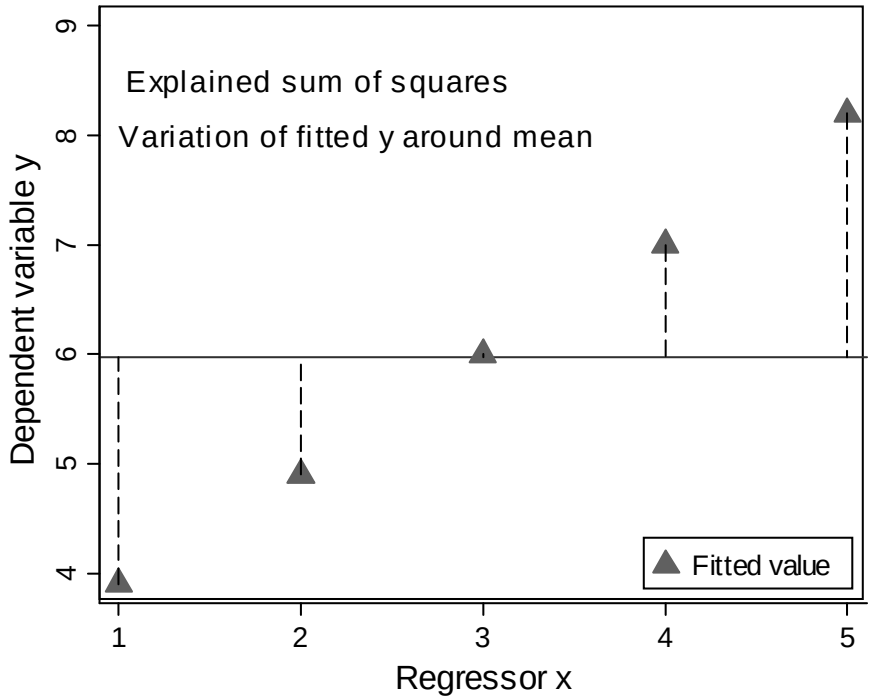
$$R^2 = \frac{Explained\ SS}{Total\ SS} = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

Example of R-squared

- Left panel is Total SS: the deviations $(y_i - \bar{y})$ for five data points.

Right panel is Explained SS: the deviations $(\widehat{y}_i - \bar{y})$ for five data points.





Computation of R-Squared

- For data in previous figure

$$\begin{aligned} \textit{Total SS} &\simeq (-1.3)^2 + (-1.3)^2 + (-1.5)^2 + 1.4^2 + 2.7^2 = 14.88 \\ \textit{Explained SS} &\simeq (-2.1)^2 + (-1.1)^2 + (0.0)^2 + 1.0^2 + 2.2^2 = 11.46 \\ R^2 &\simeq 11.46/14.88 = 0.77. \end{aligned}$$

- $R^2 = 0.77$ means 77 percent of the variation in y is explained by regression on x.

Alternative Computation of R-Squared

- **Residual sum of squares** measures variability in fitted value $\widehat{y}$ around y

$$Residual\ SS = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2,$$

- For regression including an intercept it can be shown that

$$Total\ SS = Explained\ SS + Residual\ SS$$

- As a result, R^2 can be equivalently defined as

$$R^2 = 1 - \frac{Residual\ SS}{Total\ SS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

  - So least squares maximizes R^2 as it minimizes *Residual SS.*

R-Squared and the Correlation Coefficient

- For regression of y on x that includes an intercept we have the following two results.
- R^2 equals the squared correlation coefficient between y and x

$$R^2 = r_{yx}^2$$

- R^2 equals the squared correlation coefficient between y and the fitted value $\widehat{y}$

$$R^2 = r_{y\widehat{y}}^2$$

- The second result extends to regression with additional regressors.

Interpretation of R-Squared

- Clearly $R^2 \simeq 0$ is poor fit and $R^2 \simeq 1$ is an excellent fit
  - but no rule for where R^2 becomes large enough that the fit moves from poor to good.
- The value of R^2 varies with the level of aggregation of data
  - R^2 is low for individual-level regression of earnings on education
  - R^2 is higher using aggregated data, such as state-level regression of state-average earnings on state-average schooling.
- The value of R^2 also depends on the choice of dependent variable
  - Transform y to a more symmetric distribution may increase R^2
  - Regression of levels y_t has higher R^2 than regression of changes Δy_t.
- For bivariate regression
  - use R^2 to compare models with the same dependent variable y
  - but not to compare models with different dependent variable.

Low R-Squared

- Low values of R^2 do not mean that regression analysis is without merit.
- Example: Regression of earnings on education
  - usually indicates a substantial effect of education
    - ★ e.g. one more year of education is associated with a 6% increase in annual earnings
  - yet R^2 in regressions using individual-level data is very low e.g. $R^2 = 0$.
- Explanation
  - **On average** there is a large effect of schooling on earnings.
  - At the **individual level**, however, there is considerable variability in earnings even for people with the same level of education.
  - On average, society's earnings may increase with more education, but there is great uncertainty as to whether any one given individual will necessarily see increased earnings.

Example: R-squared for House Price

- Regression output will automatically include R^2, and often $\bar{R}^2$.
- Here compute from formulas, using sums of squares that are given in "analysis of variance" table often included with regression output.

$$\begin{aligned} \textit{Explained SS} &= 24,170,725,242 \\ \textit{Residual SS} &= 14,975,101,655 \\ \textit{Total SS} &= 39,145,826,897 \\ R^2 &\simeq \frac{24,170,725,242}{39,145,826,897} = 0.6175 \\ \text{or } R^2 &\simeq 1 - \frac{14,975,101,655}{39,145,826,897} = 0.6175. \end{aligned}$$

- Thus 61.75 percent of the variation in house price is associated with variation in house size
  - this is viewed as a good fit, though still with room for improvement

5.7 Computer Output following OLS Regression

ANOVA Table

Source	SS	df	MS	F	p
Explained	2.4171×10^{10}	1	2.4171×10^{10}	43.58	0.000
Residual	1.4975×10^{10}	27	5.546×10^{8}		
Total	3.9146×10^{10}	28	1.3981×10^{9}		

Dependent Variable *Price*

Regressor	Coeff.	St. Error	t stat	p	95% C.I.	
Size	73.77	11.17	6.60	0.000	50.84	96.70
Intercept	115017	21489	5.35	0.000	70925	159110

Summary Statistics

Observations	29
F(1,27)	43.58
p-value for F	0.0000
R-squared	0.618
Adjusted R^2	0.603
St. error of reg	23551

5.8 Prediction

- For $x = x^*$ the **prediction** of y is

$$\widehat{y} = b_1 + b_2 x^*.$$

- Example: House of size 2000 square feet predicted price is \$263,000
  - $\widehat{y} = 115000 + 74 \times 2000 = 263000.$
- In-sample prediction uses the sample x_i
  - then $\widehat{y}_i$ is called the **fitted value**.
- Out-of-sample prediction
  - predictions can be poor if extrapolate to values x^* outside the sample range of x.
- Distinguish between two different uses of a prediction
  - prediction of an average outcome
    - ★ e.g. average price for a house of 2000 square feet
  - prediction of an individual outcome
    - ★ e.g. price for a particular house of 2000 square feet

Outlying Observations

- An **outlier** or **outlying observation** is one that is a relatively large distance from the bulk of the data.
- A scatter plot is a useful visual tool.
- An observation with a large value for $(x_i - \bar{x})(y_i - \bar{y})$ can have a big influence on b_2. This is the case for observations that are a long way from both $\bar{x}$ and $\bar{y}$.
- An outlier may be due to miscoded data.

5.9 Regression and Correlation

- Note: $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ appears in definitions of both b_2 and r_{xy}.
- In fact the slope coefficient

$$b_2 = r_{xy} \times \frac{s_y}{s_x}.$$

- Reason: $r_{xy} \times s_y / s_x = [s_{xy} / s_x s_y] \times s_y / s_x = s_{xy} / s_x^2 = b_2$.
- So $r_{xy} > 0 \Rightarrow b_2 > 0$ and $r_{xy} < 0 \Rightarrow b_2 < 0$.
- Also b_2 from regress $(y_i - \bar{y})/s_y$ on $(x_i - \bar{x})/s_x$ equals r_{xy}.
  - so r_{xy} measures the number of standard deviations that y changes by as x changes by one standard deviation.
  - e.g. $r_{xy} = 0.5$, $s_x = 2$ and $s_y = 10$. Then a one standard deviation change in x is associated with a 0.5 standard deviations change in y.

5.10 Causation

- The correlation coefficient always treats x and y neutrally.
- Regression does not:
  - slope b_2 from regress y on x
    $\neq$ inverse of slope c_2 from reverse regress x on y
    - ★ explained below
  - the data alone cannot tell us which direction, if any, is appropriate.
- If we estimate $y = b_1 + b_2 x$, without further information
  - can say that a one unit increase in x **is associated with** a b_2 increase in y
  - **cannot** say that a one unit increase in x causes a b_2 increase in y.

Causation (continued)

- For example: a medical study might find that alcohol consumption is associated with depression.
  - but is it alcohol consumption that causes depression
  - or is it depression that leads to alcohol consumption?
- Many examples exist where the direction of causation is questionable.
- Often it is due to a third variable that may be driving both y and x.
- For example: higher education is positively associated with higher earnings
  - but this may be due solely to unobserved innate ability that leads to both higher earnings due to higher productivity and to higher education due to ability to study more advanced material.
- Chapter 17 focuses on causality.

Reverse Regression

- **Regression of** y **on** x: $\widehat{y} = b_1 + b_2 x$
- **Reverse regression** (**of** x **on** y): $\widehat{x} = c_1 + c_2 y$.
- Then $c_1 \neq 1/b_1$!
  - In fact $c_2 = b_2 \times (s_x^2/s_y^2)$.
- For the house data
  - regression of house price on house size: $b_2 = 73.77$
  - reverse regression of house size on house price: $c_2 = 0.0084$
    - ★ whereas $1/b_2 = 1/73.77 = 0.0136 \neq 0.0084$.

5.11 Computations for Correlation and Regression

- Artificial data on number of vehicles per household (y) and household size (x)
  - $n = 5$: $(x_1, y_1) = (1, 1)$, $(x_2, y_2) = (2, 2)$, $(x_3, y_3) = (3, 2)$, $(x_4, y_4) = (4, 2)$, and $(x_5, y_5) = (5, 3)$.
- Recall want $b_2 = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n}(x_i - \bar{x})^2$.

i	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	1	1	-2	-1	2	4
2	2	2	-1	0	0	1
3	3	2	0	0	0	0
4	4	2	1	0	0	1
5	5	3	2	1	2	4
Sum	15	10	0	0	4	10
Mean	$\bar{x} = 3$	$\bar{y} = 2$				

Fitted line

- Slope, intercept and line:

$$\begin{aligned} b_2 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{4}{10} = 0.4 \\ b_1 &= \bar{y} - b_2\bar{x} = 2 - 0.4 \times 3 = 0.8 \\ \widehat{y} &= 0.8 + 0.4x. \end{aligned}$$

- Fitted values of $\widehat{y} = 0.8 + 0.4x$ for the five observations are:
  - 1.2, 1.6, 2, 2.4, and 2.8.

R-Squared

- Sum of squared residuals

$$\begin{aligned} & \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \\ = \; & (1-1.2)^2 + (2-1.6)^2 + (2-2)^2 + (2-2.4)^2 + (3-2.8)^2 \\ = \; & 0.4, \end{aligned}$$

- Total sum of squares

$$\begin{aligned} & \sum_{i=1}^{n}(y_i - \bar{y})^2 \\ = \; & (1-2)^2 + (2-2)^2 + (2-2)^2 + (2-2)^2 + (3-2)^2 \\ = \; & 2.0. \end{aligned}$$

- R-Squared

$$R^2 = 1 - \frac{0.4}{2.0} = 0.8.$$

  - 80% of the variation in number of cars is explained by household size.
- Note that the explained sum of squares is $2.0 - 0.4 = 1.6$.

Correlation Coefficient

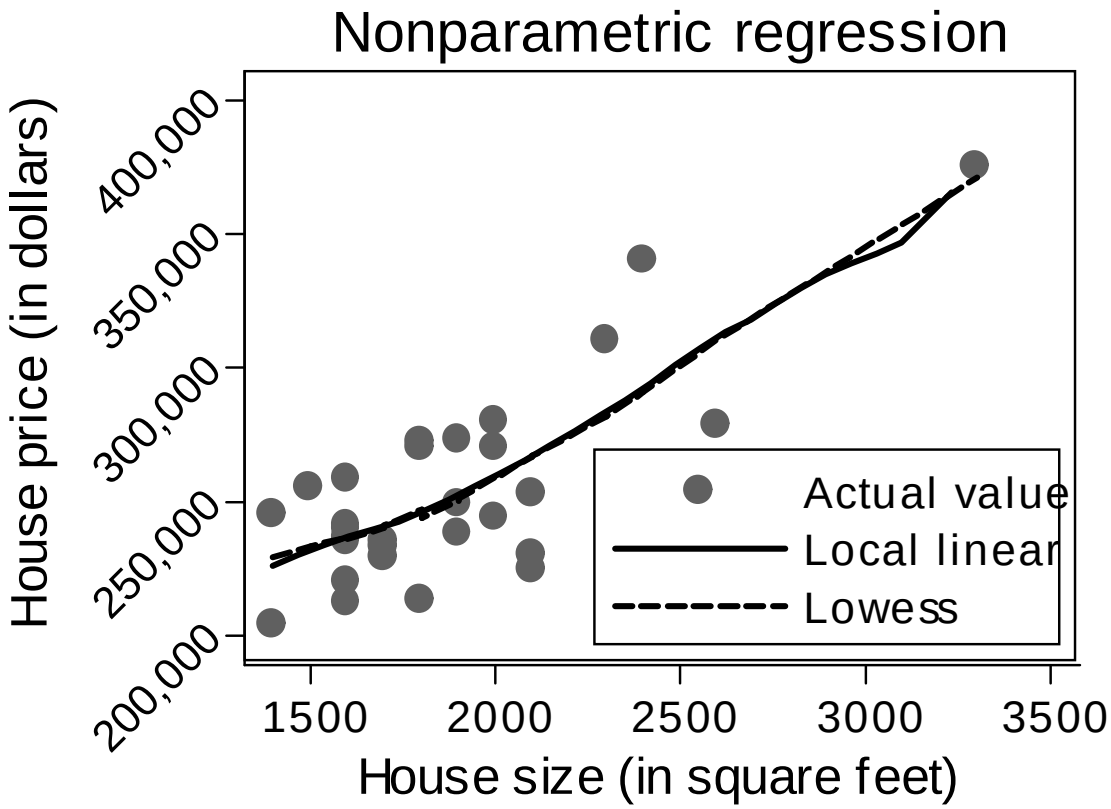
- Sample correlation coefficient

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{4}{\sqrt{10 \times 2}} = \frac{2}{\sqrt{5}} = 0.894.$$

- Close to one
  - so strong positive association between cars and household size.
- As expected, $r_{xy}^2 = (2/\sqrt{5})^2 = 4/5 = 0.8$ which equals R^2.

5.12 Nonparametric Regression

- A flexible method for exploratory data analysis
  - here the relationship appears to be linear
  - local linear and lowess are two commonly-used methods.



Key Stata Commands

```
clear
use AED_HOUSE.DTA
sort size
list price size
correlate size price
regress price size
graph twoway (scatter price size) (line price size)
display _b[_cons] + _b[size]*2000 // predict at size=2000
predict double yhat // double precision is more accurate
generate double resid = y - yhat
summarize price yhat resid // residuals sum to zero
* local linear regression
lpoly price size, degree(1) bw(300)
* lowess with default bandwidth
lowess price size, generate(ylowess)
```

Some in-class Exercises

1. Suppose we have a sample with three observations with (x, y) equal to (1,5), (2,2) and (3,2). Calculate $\sum_{i=1}^{3}(x_i - \bar{x})(y_i - \bar{y})$.
2. Variables x and y have sample variances of, respectively, 100 and 25, and their sample covariance is 8. What is the sample correlation between the two variables?
3. $\sum_{i=1}^{50}(x_i - \bar{x})^2 = 100$, $\sum_{i=1}^{50}(x_i - \bar{x})(y_i - \bar{y}) = 10$, and $\sum_{i=1}^{50}(y_i - \bar{y})^2 = 25$. Give the sample correlation between x and y.
4. For the data of the previous example, what is the slope coefficient from regression of y on an intercept and x?
5. Regression leads to fitted line $\widehat{y} = 2 + 3x$. What is the residual for observation $(x, y) = (2, 9)$?
6. OLS regression of y on x for a sample of size 52 leads to residual sum of squares 20 and total sum of squares 50. Compute the standard error of the regression.
7. For the data of the previous example, compute R^2.