

# Association and Regression

Portland State University  
USP 634 Data Analysis I  
Spring 2018

# **ASSOCIATION AND CORRELATION**

# Introduction

- Two variables are said to be associated when they vary together—that is, when one changes as the other changes.
- Association can be important evidence for causal relationships, particularly if the association is strong.

# Introduction

- If variables are associated, the score (value) of one variable can be predicted from the score of the other variable.
- The stronger the association, the more accurate the predictions.
- The “predictor” is the *independent* variable
- The variable being “predicted” is the *dependent* variable

# Association and bivariate tables

- Bivariate association can be investigated by finding answers to three questions:
  - Does an association exist?
  - How strong is the association?
  - What is the pattern and/or direction of the association?

# Association and bivariate tables

- The table shows the relationship between authoritarianism of bosses (X) and the efficiency of workers (Y) for 44 workplaces.

	Low Authoritarian	High Authoritarian	TOTAL
Low Efficiency	10	12	22
High Efficiency	17	5	22
TOTAL	27	17	44

# Is there an association?

- An association exists if the conditional distributions of one variable change across the values of the other variable.
- With bivariate tables, column percentages are the conditional distributions of  $Y$  for each value of  $X$ .
- If the column percentages change, the variables are associated.

# Association and bivariate tables

- The column % is (cell frequency / column total) \* 100.
  - $(10/27)*100 = 37.04\%$
  - $(12/17)*100 = 70.59\%$
  - $(17/27)*100 = 62.96\%$
  - $(5/17)*100 = 29.41\%$

	<b>Low authoritar.</b>	<b>High authoritar.</b>	<b>TOTAL</b>
<b>Low efficiency</b>	<b>10 (37.04%)</b>	<b>12 (70.59%)</b>	<b>22</b>
<b>High efficiency</b>	<b><u>17 (62.96%)</u></b>	<b><u>5 (29.41%)</u></b>	<b><u>22</u></b>
<b>TOTAL</b>	<b>27</b>	<b>17</b>	<b>44</b>



# Is there an association?

- The column %s show efficiency of workers (Y) by authoritarianism of supervisor (X).

	Low	High
Low	<b>37.04%</b>	<b>70.59%</b>
High	<b>62.96%</b>	<b>29.41%</b>
	<b>100%</b>	<b>100%</b>

- The column percentages change, so these variables are associated.

# How strong is the association?

- The stronger the relationship, the greater the change in column %s (or conditional distributions).
  - In weak relationships, there is little or no change in column %s.
  - In strong relationships, there is marked change in column %s.

# How strong is the association?

- One way to measure strength is to find the “maximum difference”, the biggest difference in column percentages for **any row** of the table.

Difference	Strength
Between 0 and 10%	Weak
Between 10 and 30%	Moderate
Greater than 30%	Strong

# How strong is the association?

- The maximum difference is  $70.59 - 37.04 = 33.55$ .
- This is a strong relationship.

	<b>Low</b>	<b>High</b>
<b>Low</b>	<b>37.04%</b>	<b>70.59%</b>
<b>High</b>	<b>62.96%</b>	<b>29.41%</b>
	<b>100%</b>	<b>100%</b>

# What is the pattern of the relationship?

- “Pattern” = which values of the variables go together?
- To detect, find the cell in each column which has the highest column percentage.

# What is the pattern of the relationship?

- “Low” on authoritarianism goes with “High” on efficiency.
- “High” on authoritarianism goes with “Low” on efficiency.

	<b>Low</b>	<b>High</b>
<b>Low</b>	<b>37.04 %</b>	<b>70.59 %</b>
<b>High</b>	<b>62.96 %</b>	<b>29.41 %</b>
	<b>100%</b>	<b>100%</b>

# What is the direction of the relationship?

- If *both* variables are ordinal, we can discuss *direction* as well as pattern.
- In *positive* relationships, the variables vary in the same direction.
  - As one increases, the other increases.
- In *negative* relationships, the variables vary in opposite directions.
  - As one increases, the other decreases.

# What is the direction of the relationship?

- Relationship is *negative*.
- As authoritarianism increases, efficiency decreases.
- Workplaces high in authoritarianism are low on efficiency.

	Low	High
Low	37.04 %	70.59 %
High	62.96 %	29.41 %
	100%	100%



# What is the direction of *this* relationship?

• This relationship is positive.		<b>Low</b>	<b>High</b>
• Low on X is associated with low on Y.	<b>Low</b>	<b>60%</b>	<b>30%</b>
• High on X is associated with high on Y.	<b>High</b>	<b>40%</b>	<b>70%</b>
• As X increases, Y increases.		<b>100%</b>	<b>100%</b>

# Summary

- A strong, negative relationship between authoritarianism and efficiency.
  - Consistent with the idea that authoritarian bosses cause inefficient workers (mean bosses make lazy workers).
  - **But...**
- |      | Low    | High   |
|------|--------|--------|
| Low  | 37.04% | 70.59% |
| High | 62.96% | 29.41% |
|      | 100%   | 100%   |

# Summary: Strength and direction

- ...These results may also be consistent with the idea that inefficient workers *cause* authoritarian bosses (lazy workers make mean bosses).

	Low	High
Low	37.04%	70.59%
High	62.96%	29.41%
	100%	100%

# Association vs. causation

- Association and causation are not the same things.
- Strong associations may be used as evidence of causal relationships **but** they do not prove variables are causally related.
- What else would we need to know to be sure there is a causal relationship between authoritarianism and efficiency?

# NOMINAL MEASURES OF ASSOCIATION

# Measures of association (MoAs)

- MoAs gauge strength of relationship (and do not address statistical significance).
- For nominal variables, MoAs are on 0 to 1 scale, where 0 is no relationship and 1 is strongest
- For ordinal and numeric variables, MoAs are on -1 to 1 scale,
  - where 0 is no relationship,
  - -1 is perfect negative relationship,
  - 1 perfect positive relationship

# $\chi^2$ -based MoAs: $\Phi$ [phi]

- $\chi^2$  is directly proportional to  $N$ , so can be normalized by dividing by  $N$ :  $\phi = \sqrt{\chi^2/N}$
- Provides a measure of association ranging from 0 to 1 for 2x2 tables
- $\Phi = 1 \rightarrow$  the case when the diagonally opposite cells are empty.
  - Problem with  $\Phi$  is that when Table is bigger than 2x2, upper limit  $> 1$ . Difficult to interpret.

# Spence, et al

“Approximately 64.5% of the African American respondents reported evacuating before the storm, as compared to 85.5% of the Caucasian respondents and 82.9% of other non-White evacuees,  $\chi^2 (2) = 18.67, p < .001, \phi = .143$ ”

Q: How is  $\phi$  calculated?

$$\Phi = \sqrt{\chi^2 / N} = \sqrt{18.67 / 935} = .143$$



# $\Phi$ vs. Cramer's $V$

- Cramer's  $V$ : Slightly modified  $\Phi$  suitable for larger tables:
  - The upper limit of  $\Phi$  is  $\min(r-1, c-1)$ , so divide by this term to get unity (to “normalize” to a maximum of 1).
  - Limitation: intermediate values somewhat hard to interpret because the formula is not linear. E.g., value of 0.5 not twice as strong as a value of 0.25.

$$V = \sqrt{\frac{\chi^2}{(N)(\min r - 1, c - 1)}}$$

# Limitations of $\Phi$ and Cramer's $V$

- $\Phi$  is used for 2x2 tables only. For larger tables, use  $V$ .
- $\Phi$  and  $V$  are indexes of the *strength* of the relationship *only*. They do *not* identify the pattern.
- To analyze the pattern of the relationship, see the column percentages in the bivariate table.

# Proportional reduction in error (PRE)

(Error Rate Not Knowing) – (Error Rate Knowing)

---

(Error Rate Not Knowing)

- Do your best to predict value of the dependent variable without knowing the independent variable; subtract correct predictions from total cases; this is  $E_1$  (error rate not knowing)
- Do the same using information about the independent variable (“knowing”)
- Apply the above formula

# ORDINAL MEASURES OF ASSOCIATION

# MoAs for Ordinal Variables

- Continuous ordinal variable (many possible values/scores):

- Spearman's rho 
$$r_s = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

where  $\sum D^2$  = the sum of the differences in ranks, the quantity squared

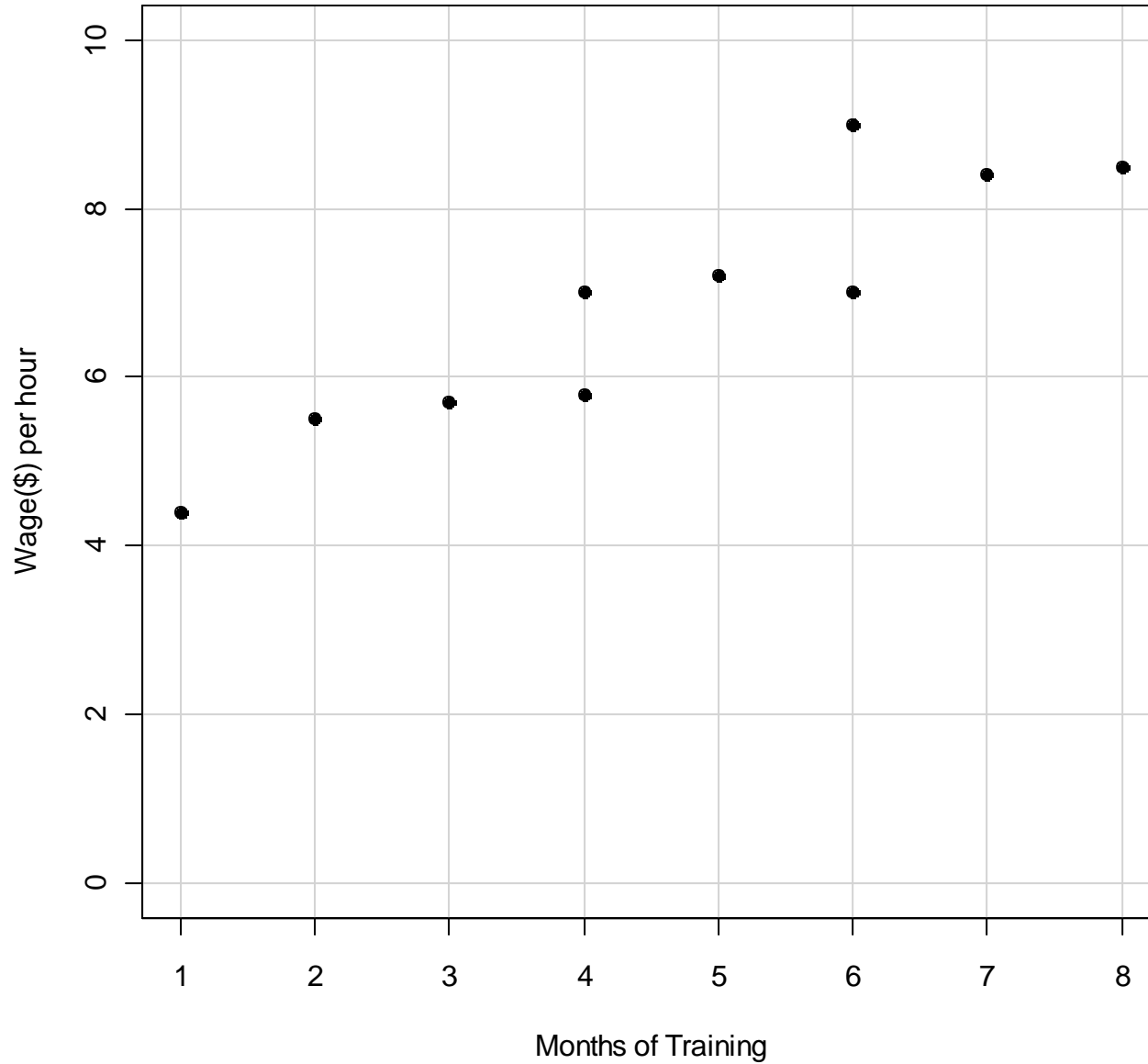
- Collapsed ordinal variable (a few values or scores):
  - Gamma (PRE)

# MEASURES OF ASSOCIATION for Numeric Variables

# Scatter plots

- Scatter plots have two dimensions:
  - The independent variable (X) is plotted along the horizontal axis (which is called “the X axis”).
  - The dependent variable (Y) is plotted along the vertical axis (which is called “the Y axis”).
- Each dot on a scatter plot is a case/an observation.
- The dot is placed at the intersection of the case’s scores on X and Y.

### Scatter Plot of Wage v.s. Months of Training



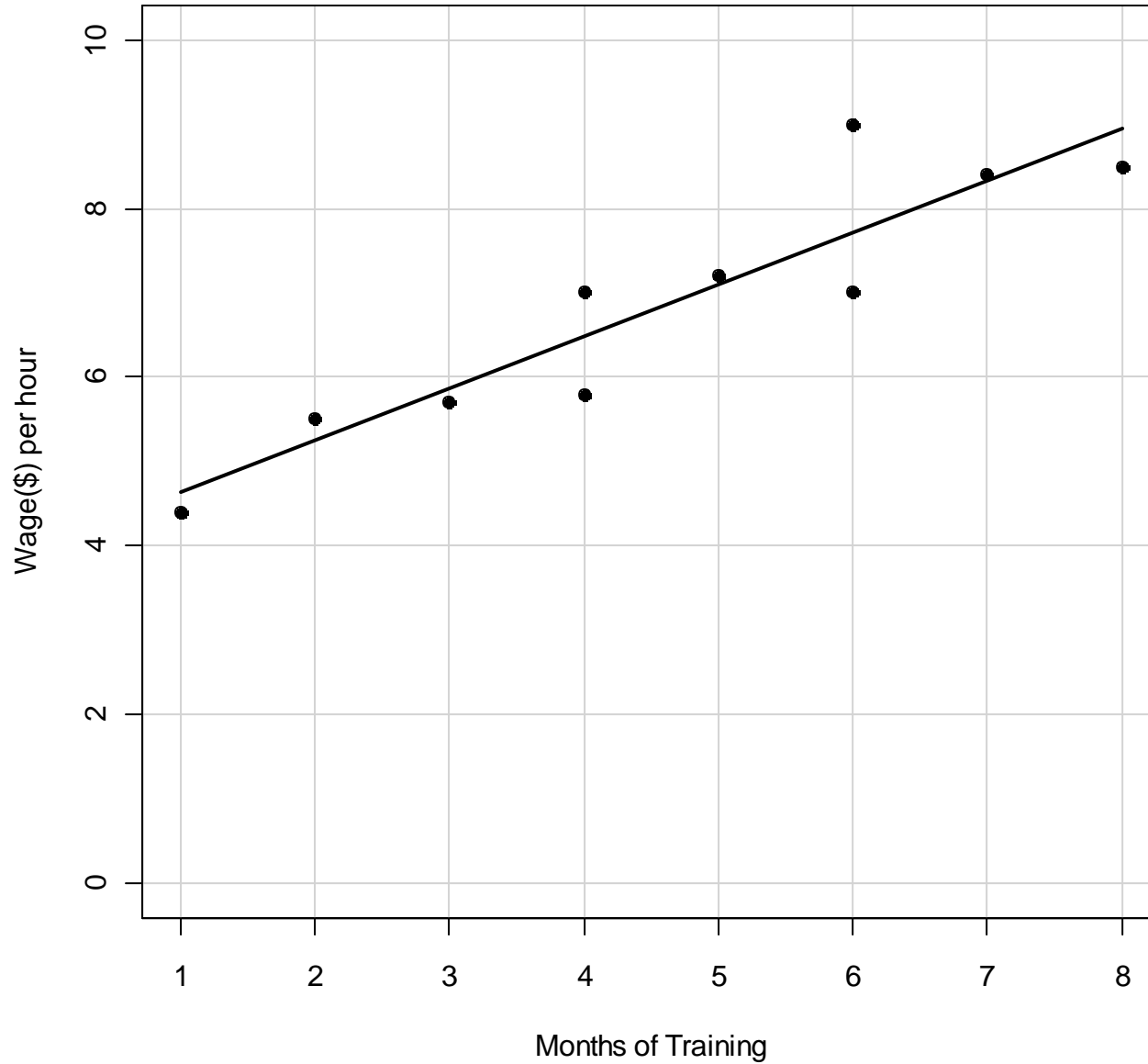
X	Y
1	4.4
2	5.5
3	5.7
4	5.8
4	7
5	7.2
6	7
6	9
7	8.4
8	8.5



# Scatter plot & regression line

- Inspection of the scatter plot should always be the first step in assessing the association between two numeric variables
- Regression line is a single straight line that comes “as close as possible” to all data points, which indicates **strength** and **direction** of the relationship

### Scatter Plot of Wage v.s. Months of Training



<u>X</u>	<u>Y</u>
1	4.4
2	5.5
3	5.7
4	5.8
4	7
5	7.2
6	7
6	9
7	8.4
8	8.5

# Regression line: Strength and direction

- Strength of association
  - The greater the extent to which dots are clustered around the regression line, the stronger the relationship
- Direction of association
  - Positive: regression line rises left to right.
  - Negative: regression line falls left to right.
- Slope of regression line
  - Steeper slope implies larger “effect” —but caution: this partly an artifact of variable *units* and outliers

# How do we measure the association of X and Y?

- Use a calculated regression line, if linear relationship is appropriate
- Another way to measure the extent of clustering around the regression line is to use Pearson's  $r$  or  $R^2$ . These measures can be tested for statistical significance.

# Pearson's r

- AKA Pearson Product-Moment **Correlation**
- Pearson's r is a measure of association for numeric variables:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \sqrt{\sum(Y_i - \bar{Y})^2}}$$

- Ranges from -1 to 1:
  - 0 indicates no relationship,
  - -1 a perfect negative relationship
  - 1 a perfect positive relationship
- Limitation: No direct interpretation of intermediate values

# Correlation: Pearson's r

- $$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \sqrt{\sum(Y_i - \bar{Y})^2}} = \frac{\sum X_i Y_i - N \bar{X} \bar{Y}}{\sqrt{\sum X_i^2 - N \bar{X}^2} \sqrt{\sum Y_i^2 - N \bar{Y}^2}}$$

- $$r = \frac{342.5 - 10 * 4.60 * 6.85}{\sqrt{256 - 10 * 4.60^2} \sqrt{489.4 - 10 * 6.85^2}} = 0.916$$

- R code: `cor(X, Y)`

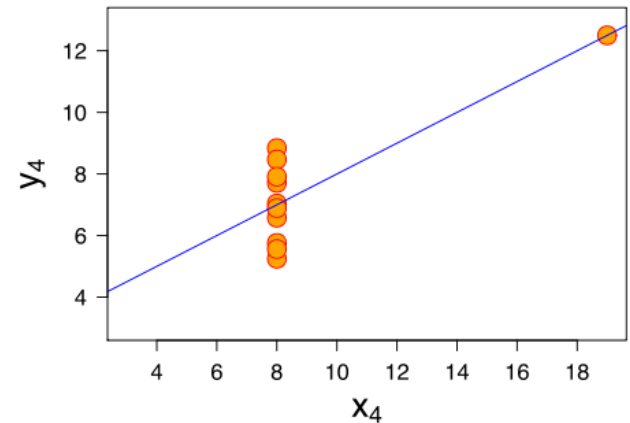
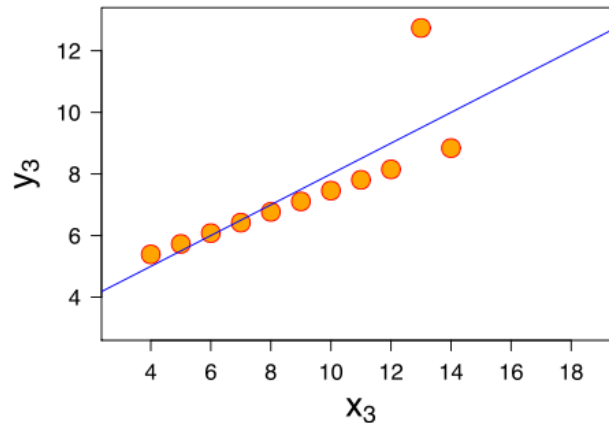
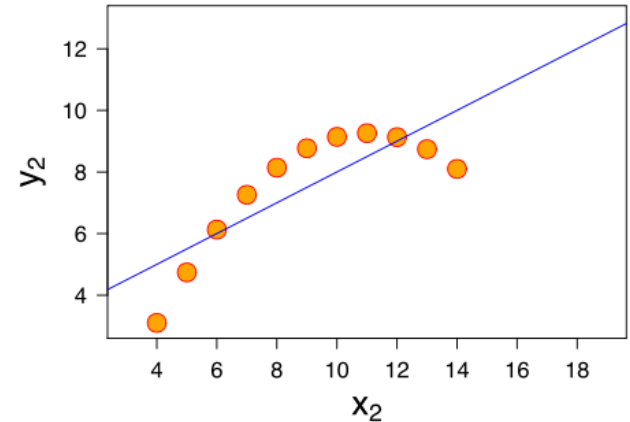
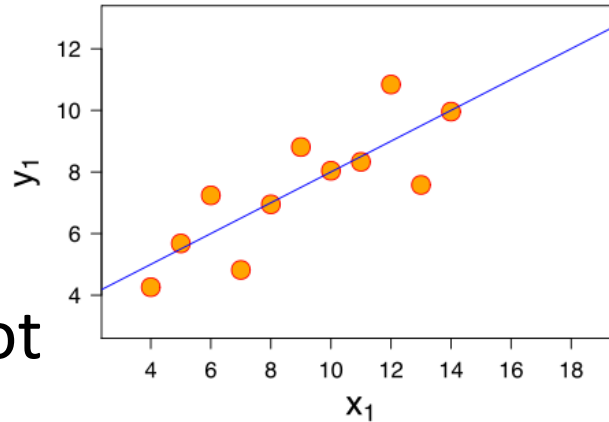


$N = 10$   
 $\sum X_i = 46$   
 $\sum X_i^2 = 256$   
 $\sum Y_i = 68.5$   
 $\sum Y_i^2 = 489.4$   
 $\sum X_i Y_i = 342.5$   
 $\bar{X} = 4.60$   
 $\bar{Y} = 6.85$

<u>X</u>	<u>Y</u>
1	4.4
2	5.5
3	5.7
4	5.8
4	7
5	7.2
6	7
6	9
7	8.4
8	8.5

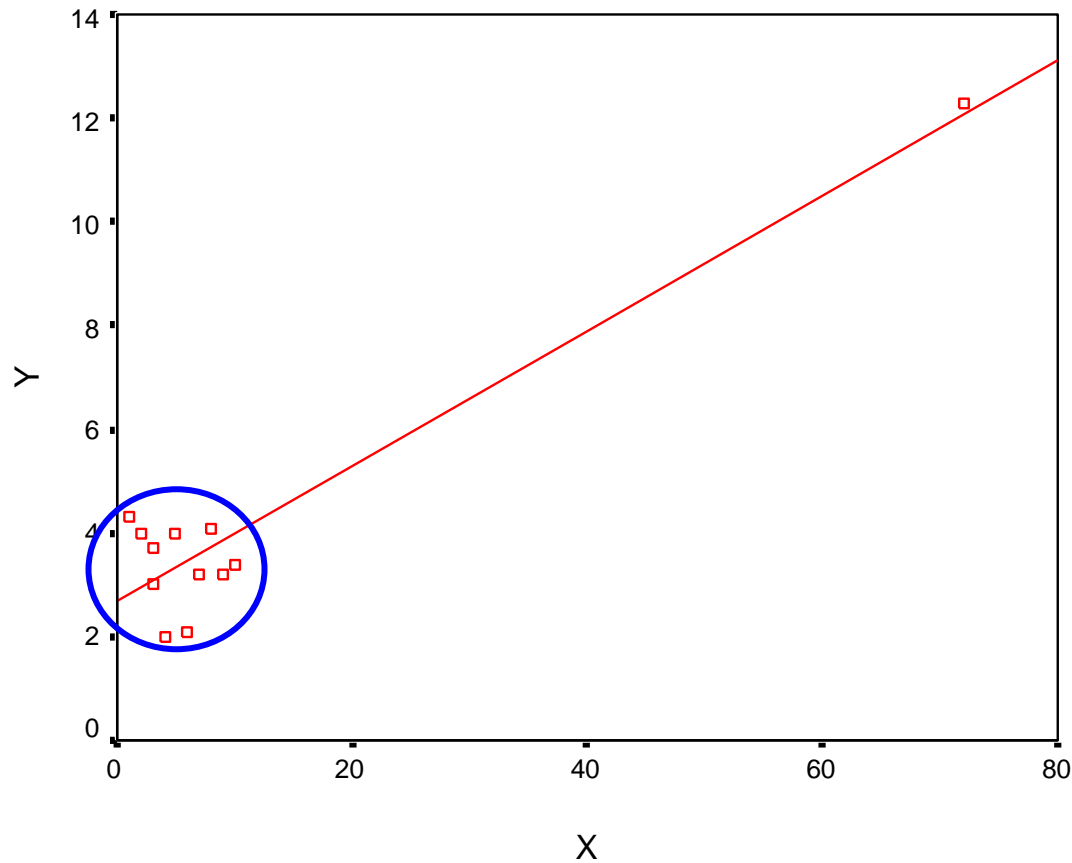
# Correlation

- Assume a linear relationship
- Sensitive to outliers
- Always look at scatter plot, not just  $r$  statistic.



Four sets of data with the same correlation of 0.816 <sup>50</sup>

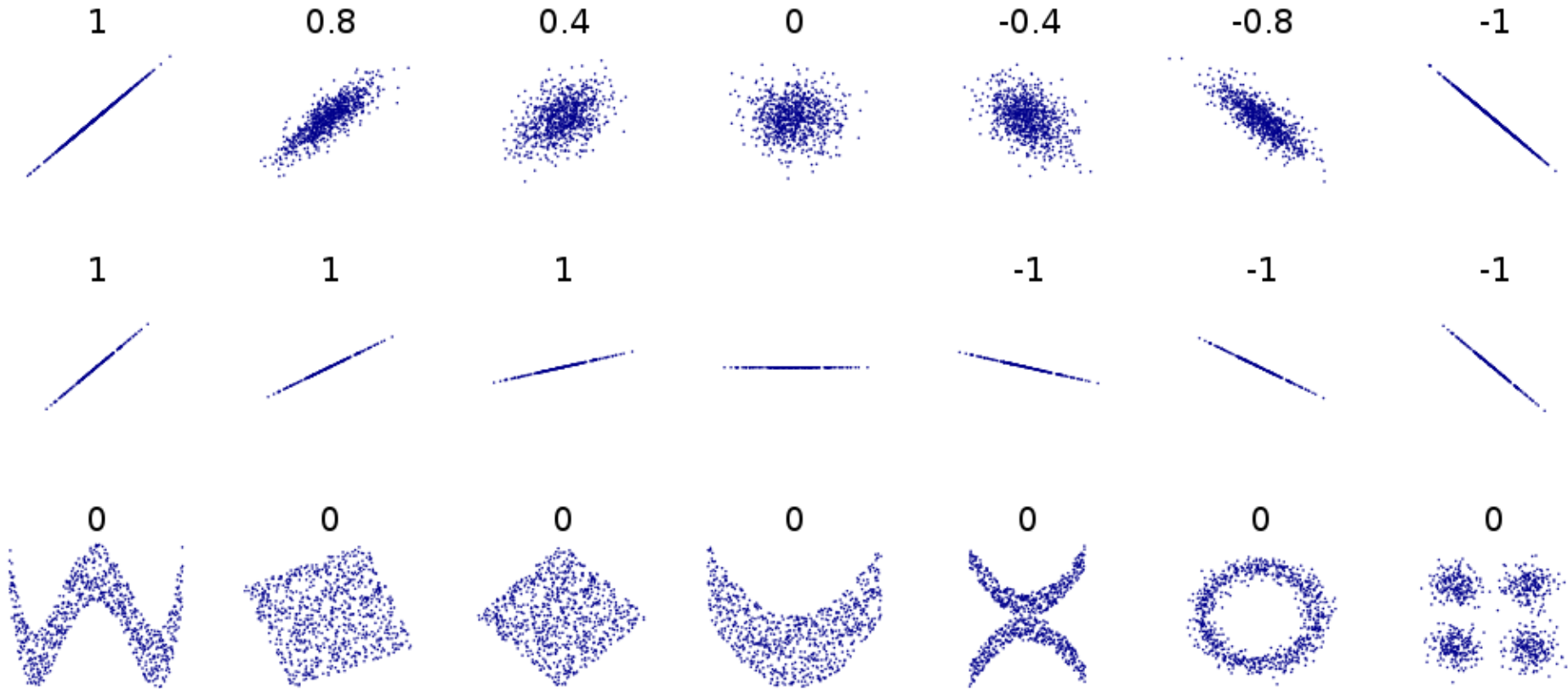
# Correlation



**Outlier:** high  $r$ ; but for most data points, no relationship.



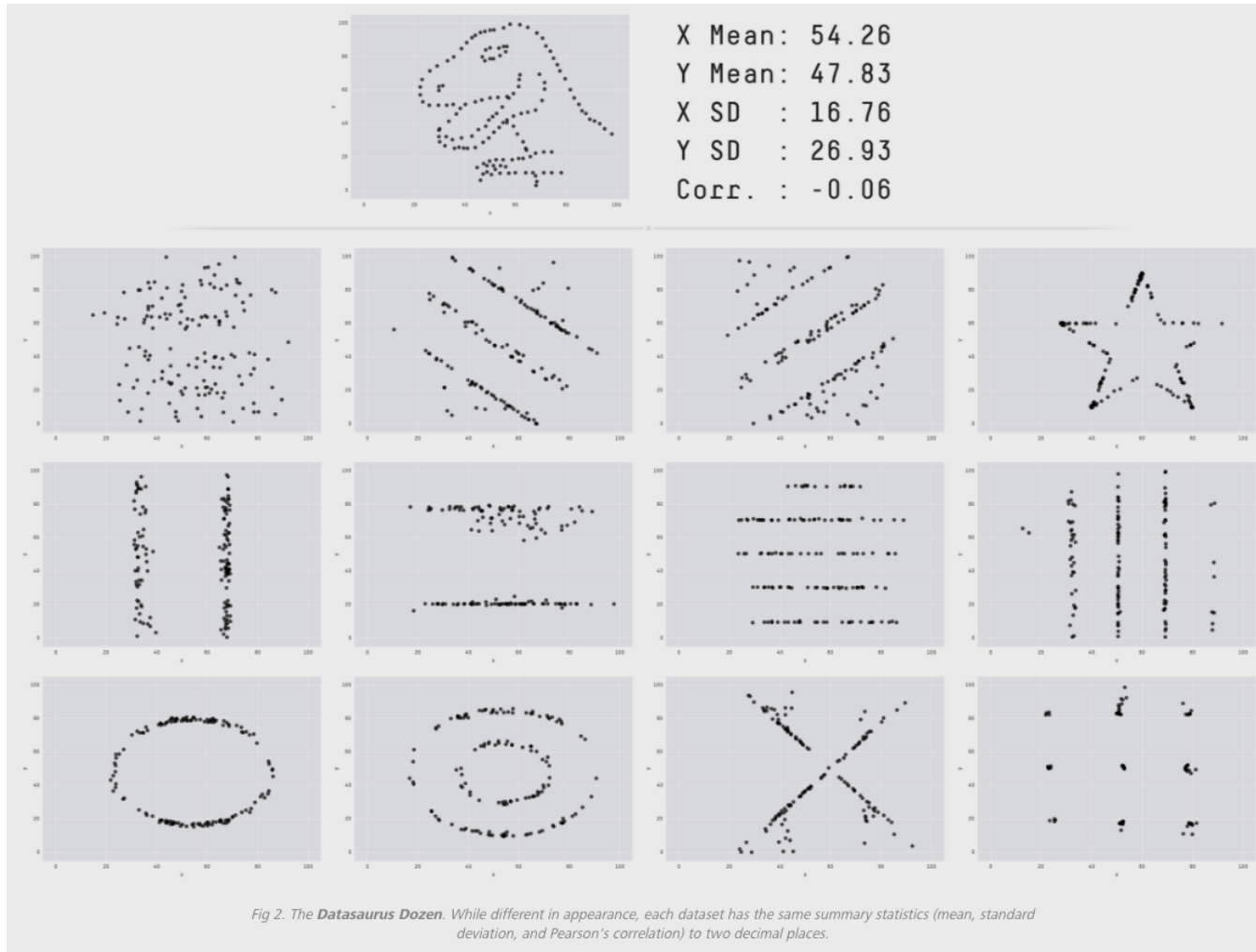
# Correlation



## Subjective Strength

0 - .30 Weak  
.30 - .60 Moderate  
> .60 Strong

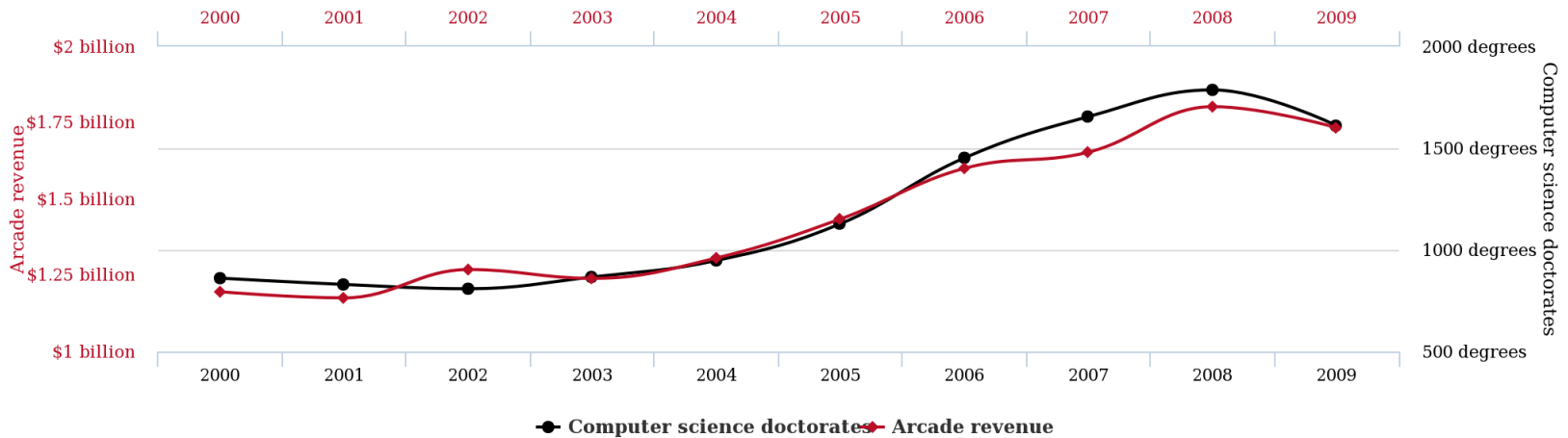
# Always plot your data!



Source: <https://www.autodeskresearch.com/publications/samestats>

# Spurious Correlation

**Total revenue generated by arcades**  
correlates with  
**Computer science doctorates awarded in the US**



tylervigen.com

<http://www.tylervigen.com/spurious-correlations>

# Coefficient of Determination ( $R^2$ )

- By squaring  $r$ , we obtain a PRE measure called the **coefficient of determination** ( $R^2$ )
- Can be interpreted as the proportion of variation in dependent variable ( $Y$ ) explained by independent variable ( $X$ ):

$$R^2 = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2} = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

- *Our example:*  $R^2 = (.916)^2 = .84$  (84% of variation in hourly wages is accounted for by months of training).

# Testing Pearson's $r$ for significance

1. Hypothesis:

1.  $H_0: r = 0$

2.  $H_1: r \neq 0$

2. Assumptions: 1) random sampling, 2) normal distributions, 3) a linear relationship, and 4) equal variance of  $y$  for all values of  $x$  ("homoskedasticity")

3. Sampling distribution is student's  $t$ , with d.f.= $N-2=8$ :  $t_{\alpha=0.05/2} = \pm 2.306$

4. Test Statistic is  $t(\text{obtained}) = r \sqrt{\frac{N-2}{1-r^2}}$

$$t(\text{obtained}) = .916 \sqrt{\frac{10-2}{1-.916^2}} = 6.45 > t_{\alpha=0.05/2} : \text{Reject } H_0$$

5. Conclusion

# Correlation Matrix

- A **correlation matrix** is a table that shows the relationships between all possible pairs of variables
- Using the matrix below:
  - What is the correlation between Birth Rate and Infant Mortality Rate?
  - Of all the variables correlated with Infant Mortality Rate, which has the strongest relationship? The weakest?

**A Correlation Matrix Showing the Interrelationships of Four Variables for 74 Nations**

	1	2	3	4
	Birth Rate	Infant Mortality Rate	Life Expectancy	Percent Urban
1 Birth Rate	1.00	0.88	-0.84	-0.66
2 Infant Mortality Rate	0.88	1.00	-0.89	-0.71
3 Life Expectancy	-0.84	-0.89	1.00	0.74
4 Percent Urban	-0.66	-0.71	0.74	1.00

# A lot more measures out there

- Darlington, Richard. **An Outline for Choosing among 19 Measures of Association** (for categorical variables):

<http://www3.psych.cornell.edu/Darlington/crosstab/TABLE0.HTM>

- Correlation (for numeric variables)

[http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5\\_Correlation-Regression/R5\\_Correlation-Regression3.html](http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression3.html)

# Measures of Bivariate Associations

		Independent Variable		
		Nominal 2 Groups	Ordinal	Numeric
Dependent Variable	Nominal	Phi, Cramer's V; Lambda;	Phi, Cramer's V; Lambda;	--
	Ordinal	Phi, Cramer's V; Lambda;	Gamma; Kendall's Tau; Somer's d	--
	Numeric	* Mann-Whitney U * Runs	--	Spearman's rho Pearson's r Kendall's Tau

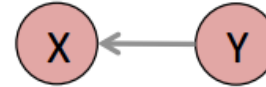


# Linear Regression

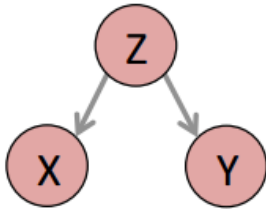
# How Correlation Happens



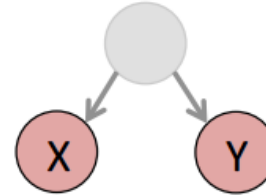
X causes Y



Y causes X



Z causes X and Y



hidden variable causes X and Y



random chance!

# Linear Regression and Regression Analysis

- Used to estimate a relationship between a numeric dependent variable & one or more independent variables (*numeric or categorical*).
- Used to:
  - **Build theory**: tests hypotheses; controls for other independent variables; rule out spurious relationships
  - **Forecast**: Can *predict* outcomes using estimated equations

# Read regression output

```
> summary(m <- lm(mpg~wt, data=mtcars))
```

```
Call:
```

```
lm(formula = mpg ~ wt, data = mtcars)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776   19.858 < 2e-16 ***
wt           -5.3445     0.5591   -9.559 1.29e-10 ***
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Testing  $H_0: b = 0$

```
Residual standard error: 3.046 on 30 degrees of freedom
```

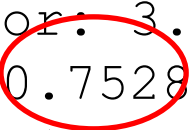
```
Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
```

```
F-statistic: 91.38 on 1 and 30 DF, p-value: 1.294e-10
```

Hypothesis Testing  $H_0: r^2 = 0$

# What percent of the variation in mpg can be explained by the variation in wt?

```
Residual standard error: 3.046 on 30 degrees of freedom  
Multiple R-squared: 0.7528, Adjusted R-squared: 0.7446  
F-statistic: 91.38 on 1 and 30 DF, p-value: 1.294e-10
```



Regression line does an 75% better job of predicting mpg than the mean value of mpg

```
> cor(mtcars$mpg, mtcars$wt)  
-0.868
```

Pearson's  $r$  for these 2 variable is -0.867, and its squared value, coefficient of determination, is  $-0.868^2 = 0.753$

# Two Main Significance Tests in a Linear Regression Model

1. F test of the equation ( $H_0: r^2 = 0$ ) using ANOVA F-test

$$F \text{ statistic} = \frac{\sum(\hat{Y}_i - \bar{Y})^2 / df_1}{\sum(Y_i - \hat{Y}_i)^2 / df_2} = \frac{R^2(N-2)}{1-R^2}$$

2.  $t$  test of coefficient:  $H_0: b = 0$

$$t \text{ statistics} = \frac{b-0}{SE(b)}$$

In a bivariate regression (regression with one independent variable) analysis, they're equivalent.

# Equivalency between Regression and ANOVA

```
> summary(anova <- aov(mpg~vs,
  data=mtcars))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
vs	1	496.5	496.5	23.66	3.42e-05 ***
Residuals	30	629.5	21.0		

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05  
'.' 0.1 ' ' 1

```
> summary(m <- lm(mpg~vs, data=mtcars))
```

```
Call:
lm(formula = mpg ~ vs, data = mtcars)
Residuals:
    Min       1Q   Median       3Q      Max
-6.757 -3.082 -1.267  2.828  9.383

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   16.617      1.080   15.390 8.85e-16 ***
vs              7.940      1.632    4.864 3.42e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
```

```
Residual standard error: 4.581 on 30 degrees of
freedom
Multiple R-squared:  0.4409,    Adjusted R-squared:
0.4223
F-statistic: 23.66 on 1 and 30 DF,  p-value: 3.416e-05
```

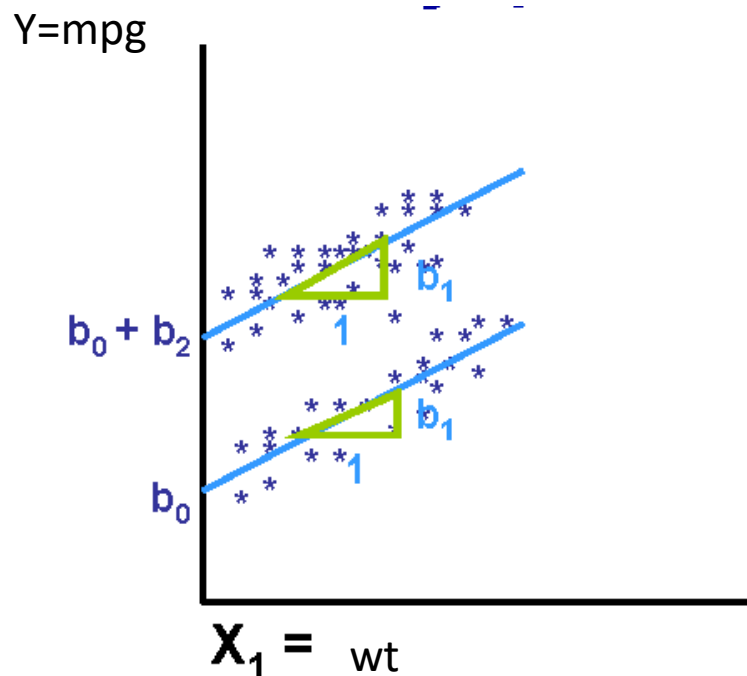
# “Dummy” variables

- A binomial variable taking values 1 and 0
- The coefficient indicates the effect in being in one category (assigned value “1”) in comparison to the effect of being in another category (assigned value “0”)
- You can create binomial variables from ordinal variables or from nominal/categorical variables



# Regular or “fixed effect” dummy variables

- $Y = b_0 + b_1X_1$ 
  - Y: mpg
  - $X_1$ : wt
- Add  $X_2$ , which is a dummy variable equal to 1 if a cars has V engine
- $Y = b_0 + b_1X_1 + b_2X_2$



```
> summary(lm(mpg~wt+vs, data=mtcars))
```

```
Call:
```

```
lm(formula = mpg ~ wt + vs, data = mtcars)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-3.7071	-2.4415	-0.3129	1.4319	6.0156

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	33.0042	2.3554	14.012	1.92e-14	***
wt	-4.4428	0.6134	-7.243	5.63e-08	***
vs	3.1544	1.1907	2.649	0.0129	*

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.78 on 29 degrees of freedom
```

```
Multiple R-squared:  0.801, Adjusted R-squared:  0.7873
```

```
F-statistic: 58.36 on 2 and 29 DF,  p-value: 6.818e-11
```

# Categorical Variable where k categories > 2 → Multiple Dummies

- Create k-1 dummies (where k = # categories)
- **gear** (k=3): 3; 4; 5
- **2 Dummies:** G4 = 4 gears (0=no; 1=yes)

G5 = 5 gears (0=no; 1=yes)

*Note:* 3 gear is suppressed (as the reference group)

$$\hat{Y} = b_0 + b_1X_1 + b_2G4 + b_3G5 \quad Y = \text{mpg}; X_1 = \text{wt};$$

G4 = 4 gears ; G5 = 5 gears

If 3 gear:  $\hat{Y} = b_0 + b_1X_1$

If G4=1:  $\hat{Y} = (b_0 + b_2) + b_1X_1$

If G5=1:  $\hat{Y} = (b_0 + b_3) + b_1X_1$

```

> summary(lm(mpg~wt+as.factor(gear), data=mtcars))

Call:
lm(formula = mpg ~ wt + as.factor(gear), data = mtcars)

Residuals:
    Min       1Q   Median       3Q      Max
-3.517 -2.358 -0.355  1.850  5.821

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      35.2156     2.8690  12.274 8.72e-13 ***
wt                -4.9090     0.7112  -6.902 1.68e-07 ***
as.factor(gear)4    2.1631     1.4485   1.493  0.147
as.factor(gear)5   -0.9121     1.7519  -0.521  0.607
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.915 on 28 degrees of freedom
Multiple R-squared:  0.7887,    Adjusted R-squared:  0.766
F-statistic: 34.83 on 3 and 28 DF,  p-value: 1.375e-09

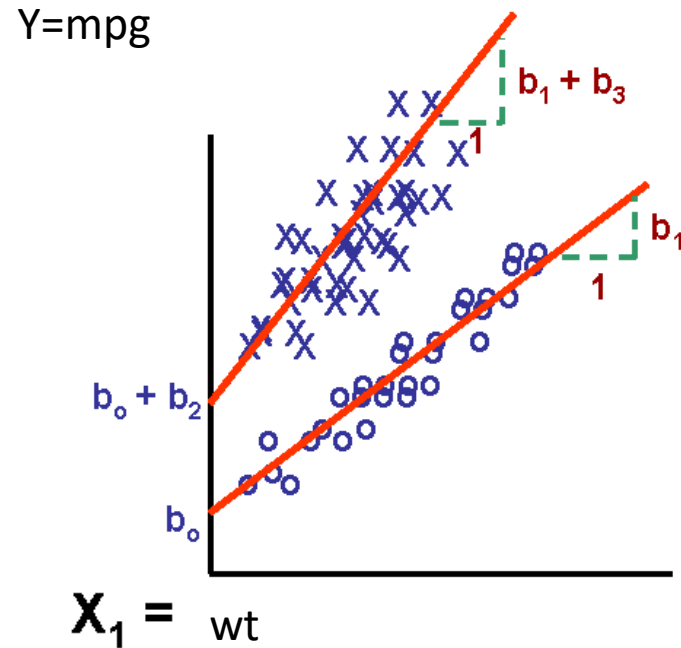
```

# Interactive dummy variables

- Take a dummy variable and multiply it by some other variable (sometimes a continuous variable, sometimes another dummy variable) to create a new variable;
- The “interaction” is the marginal difference in slope or effect for the subgroup represented by dummy value “1”

# Interactive dummy variables

- Perhaps the *effect* of wt on mpg is different in V engine cars vs S engine cars
- Create new variable
  - $X_2 * X_1$  (let's call that  $X_3$ )
- $Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$



```
> summary(lm(mpg~wt*vs, data=mtcars))
```

```
Call:
```

```
lm(formula = mpg ~ wt * vs, data = mtcars)
```

```
Residuals:
```

	Min	1Q	Median	3Q	Max
	-3.9950	-1.7881	-0.3423	1.2935	5.2061

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	29.5314	2.6221	11.263	6.55e-12	***
wt	-3.5013	0.6915	-5.063	2.33e-05	***
vs	11.7667	3.7638	3.126	0.0041	**
wt:vs	-2.9097	1.2157	-2.393	0.0236	*

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.578 on 28 degrees of freedom
```

```
Multiple R-squared:  0.8348,      Adjusted R-squared:  0.8171
```

```
F-statistic: 47.16 on 3 and 28 DF,  p-value: 4.497e-11
```

# “Art & Science” of Model Building

- ***Model Building***: What variables to include & in what form; match, refine, modify, build theories.
  - High explanatory power (high  $R^2$ )
  - Adhere to principle of parsimony
  - Pass “reasonableness” test

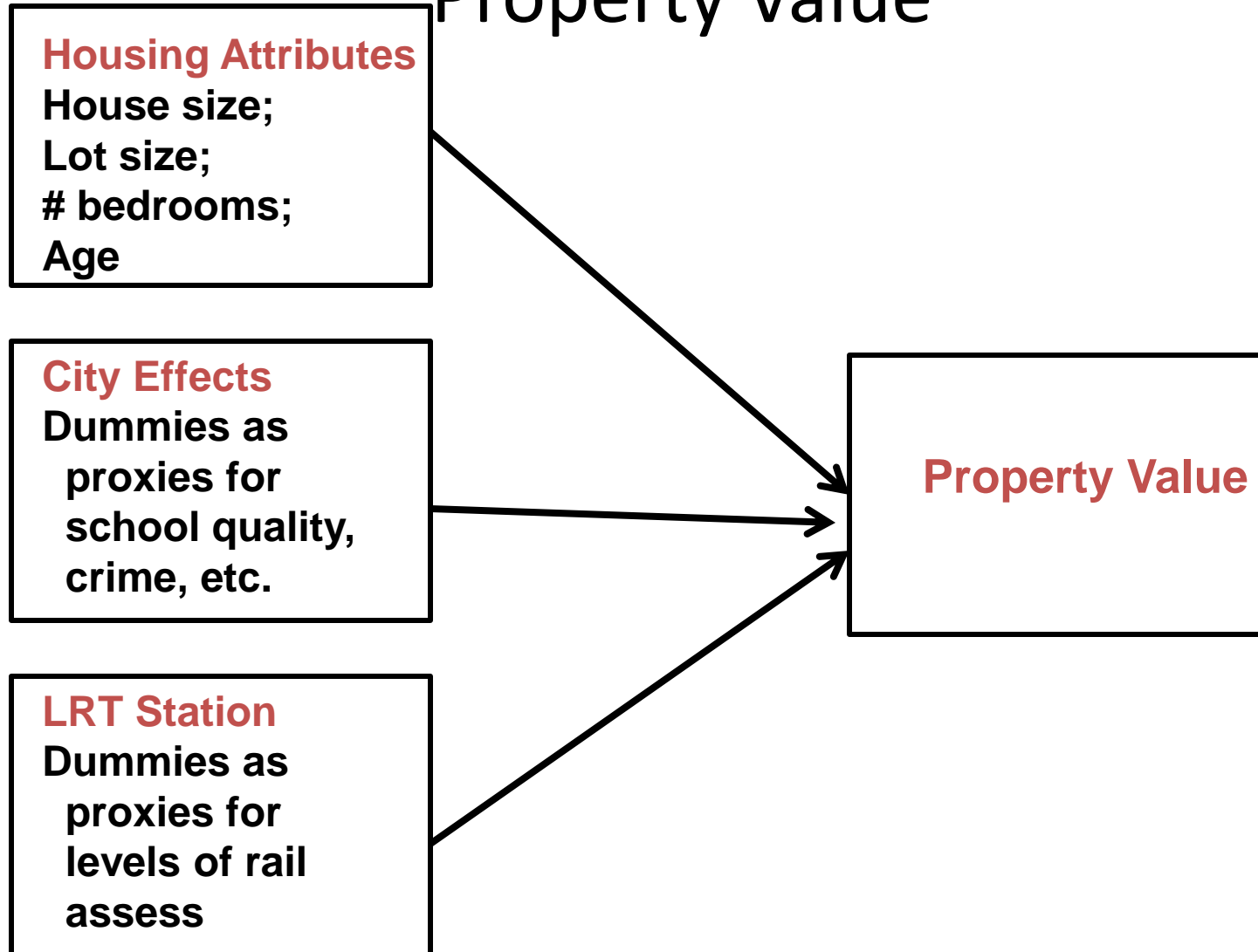


# Steps of Model Building

- 1) Formulate Research Question:** Draw path diagram representing theory
- 2) Plot scatterplots** (check for non-linearity; violation of assumptions); generate correlation matrices.
- 3) Decide variables to include into model:** *Exploratory technique:* Stepwise regression
- 4) Diagnostics:** generate residual plots of final model
- 5) Conduct “reasonableness” test** (signs intuitive?)
- 6) How do results match with initial theories/ postulates?**  
Revise theories?
- 7) What are planning/policy implications of study findings?** Forecasts? Sensitivity Tests?

# Path Diagram

## Hedonic Price Model: Impact of Light Rail on Property Value



# Impacts of MAX stations on property value

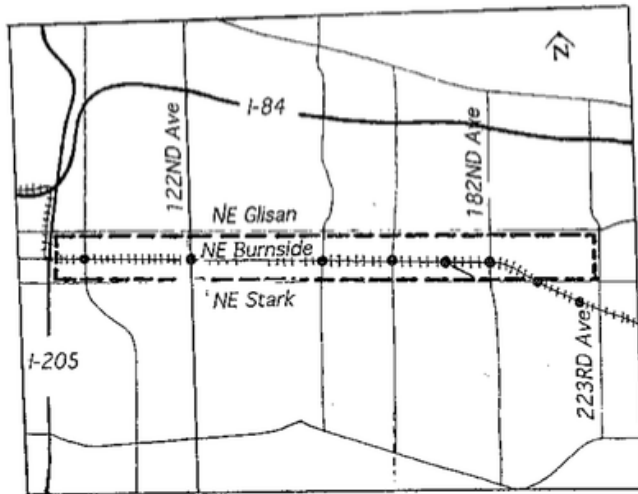


FIGURE 2 The study area.

TABLE 1 Results of Linear Regression of All Homes

Variable	Coefficient	T-score
Distance from nearest station (1=within 500 m. <sup>1</sup> , 0=further)	4324	2.49*
Lot size in sq. meters <sup>2</sup>	3.98	4.19**
House size in sq. meters	210.35	6.67**
Presence of Basement (1=Yes, 0=No)	6330	3.75**
Number of bedrooms	3398	2.24*
Age of house in years	-384	-6.32**
Single family zoning (1=Yes, 0=No)	6661	3.46**
Located in Portland (1=Yes, 0=No)	4476	2.40*
Located in Multnomah County (1=Yes, 0=No)	6583	3.62**
Constant	16919	
Number of cases	235	
Coefficient of Determination (R <sup>2</sup> )	.631	
Standard error of estimate	11018	
F-Ratio	42.66**	

<sup>1</sup> 1 meter = 3.28 feet.

<sup>2</sup> 1 sq. meter = 10.76 sq. feet.

\* Significant at the 0.05 level (two-tailed test).

\*\* Significant at the .005 level (two-tailed test).

# Diagnose Ordinary Least Squares (OLS) Estimate

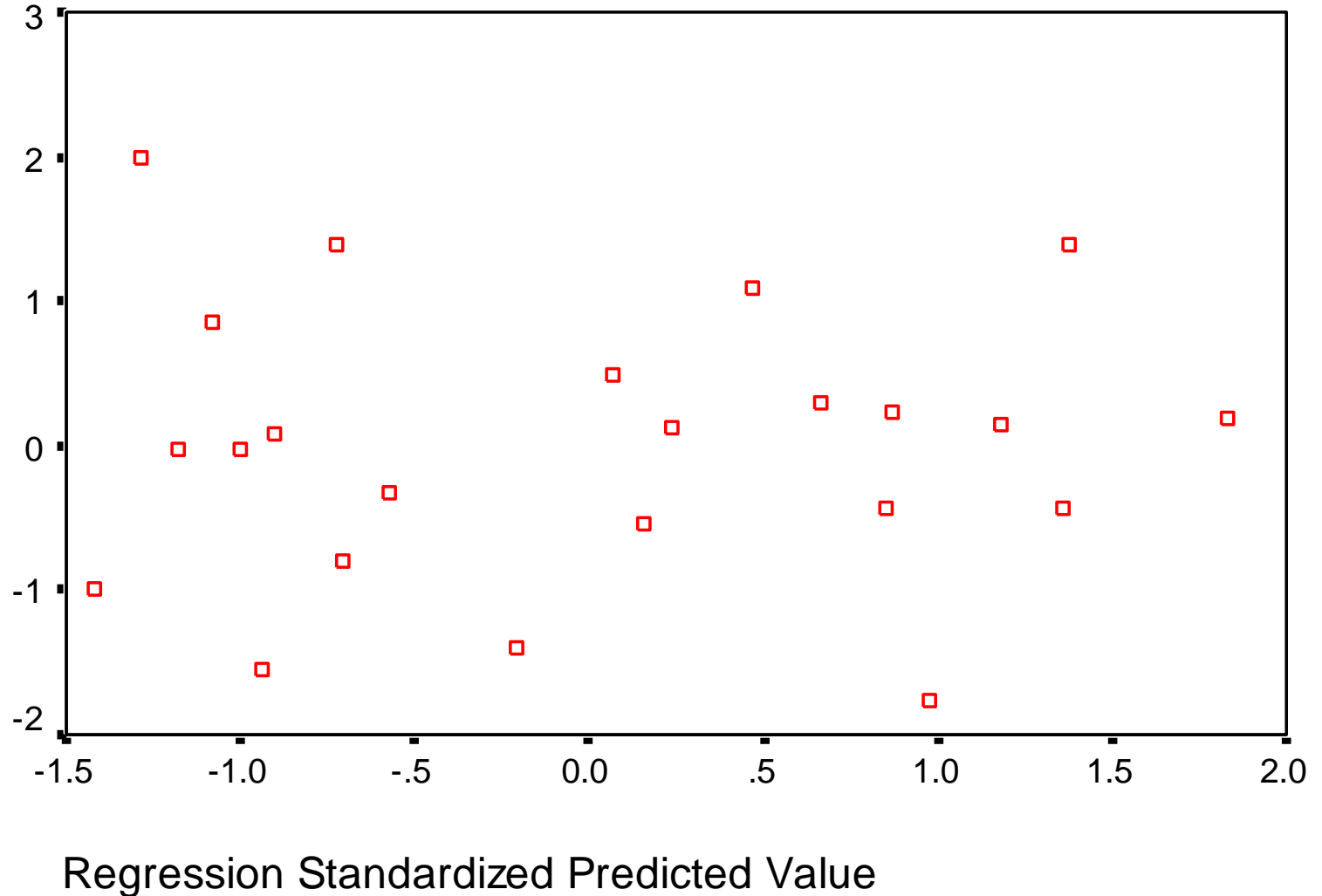
- OLS is the linear regression procedure for estimating  $a$  and  $b$  (aka  $\alpha$  and  $\beta$ )
- OLS produces **best** (efficient) **linear unbiased estimates** (BLUE) of  $a$  and  $b$  under the following assumptions of the error term (residual,  $e_i = Y_i - \hat{Y}_i$ ):
  - Equal variance (shape)
  - Uncorrelated to predicted values and ind. variable
  - *Normally distributed*

# Diagnostic Plots

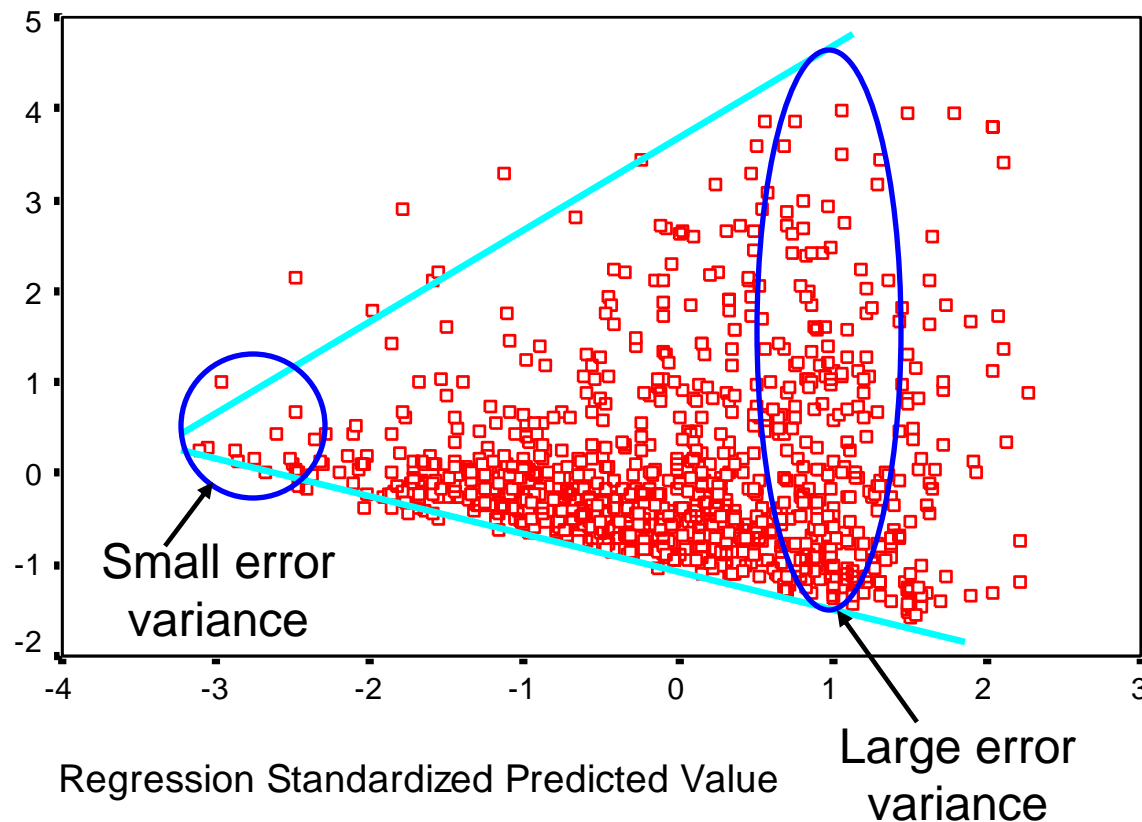
- Used as visual diagnostics to examine whether error term assumptions are met
  - Residual Plots:
    - $e_i$  versus  $\hat{Y}_i$
    - $e_i$  versus  $X_i$
    - To examine residuals, take out measurement units by standardizing
  - Normal Q-Q plot

# Residual plot

Want a Random Pattern: Suggests error term assumptions are met



Suggests violation of assumption of Equal Error Variance (homoscedasticity)

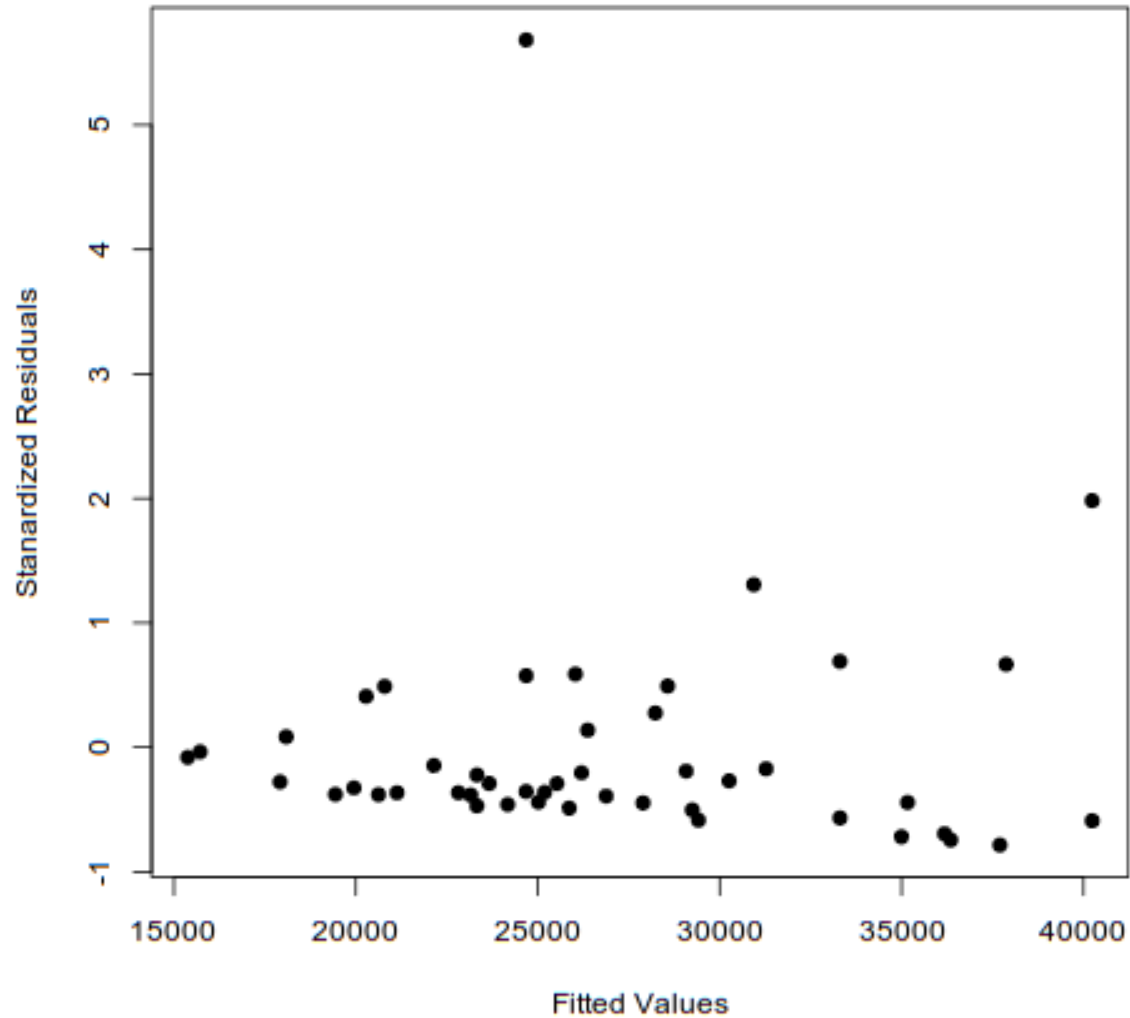


**Problem: heteroscedasticity...**

→ use alternative estimation approach

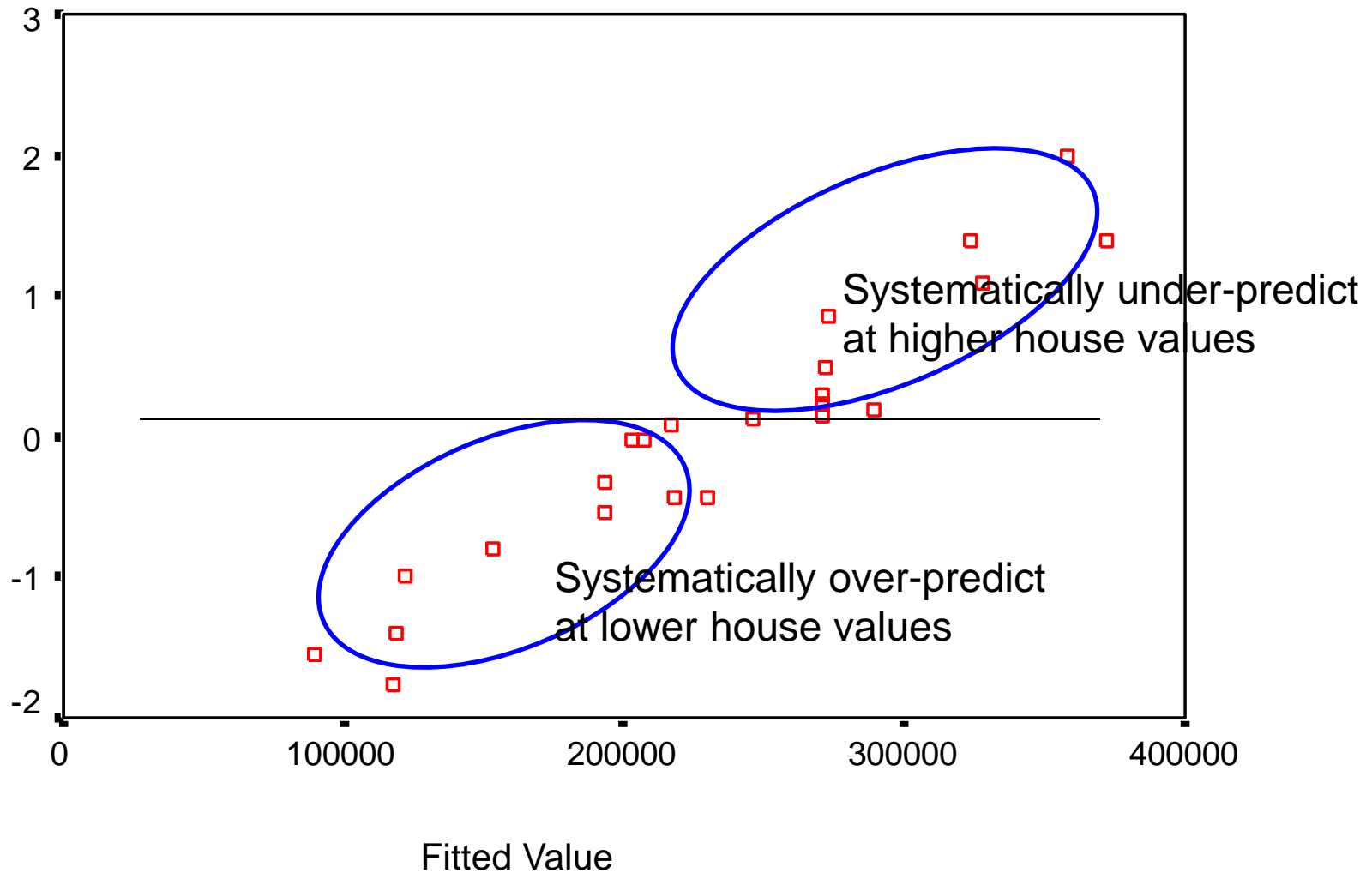
# Identifies Potential Outliers

Standardized Residuals vs Fitted Values



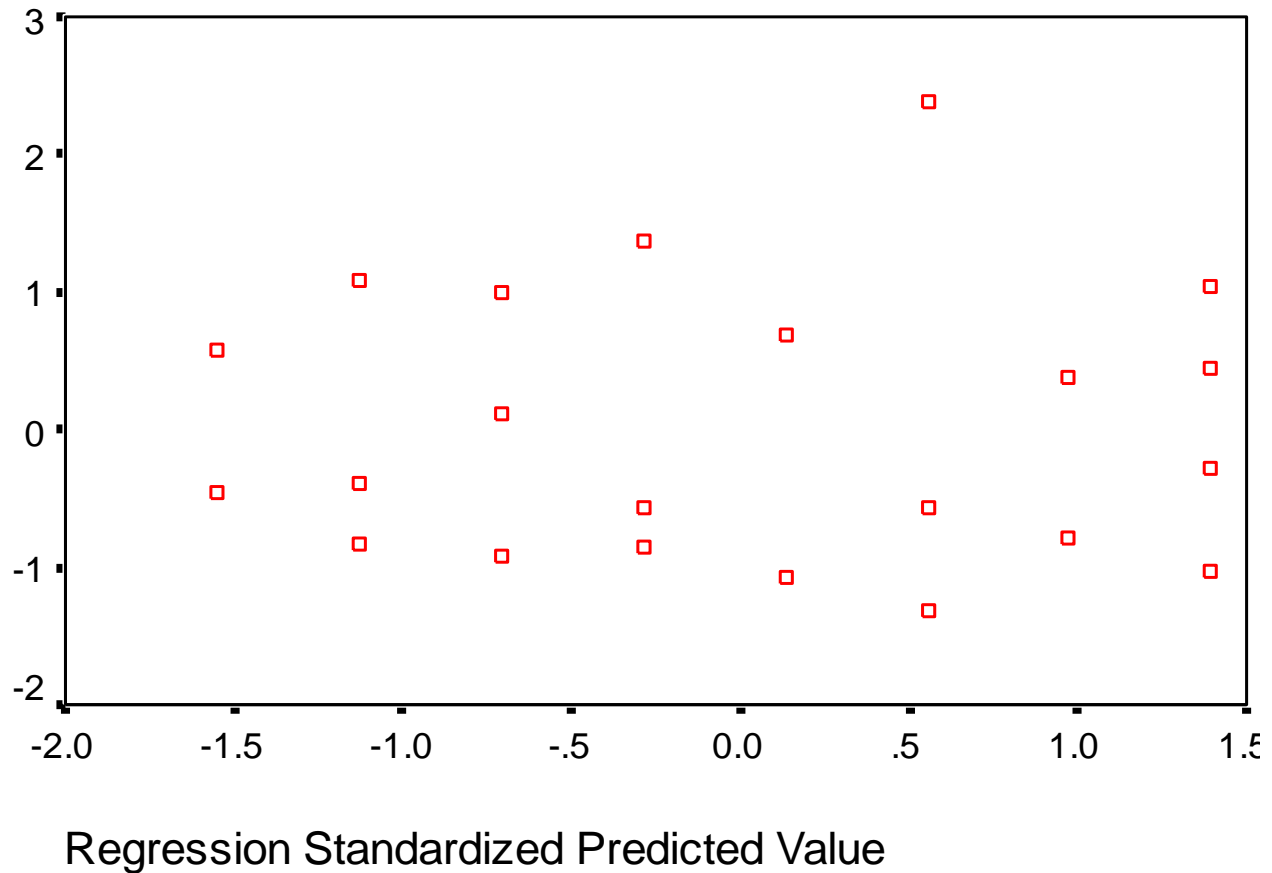


# Sign of an under-specified model: needs multiple regression

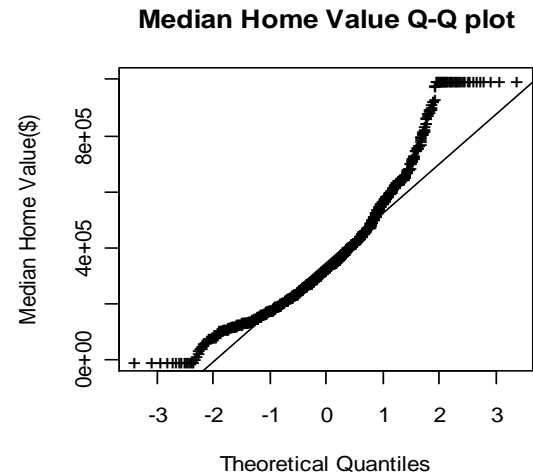
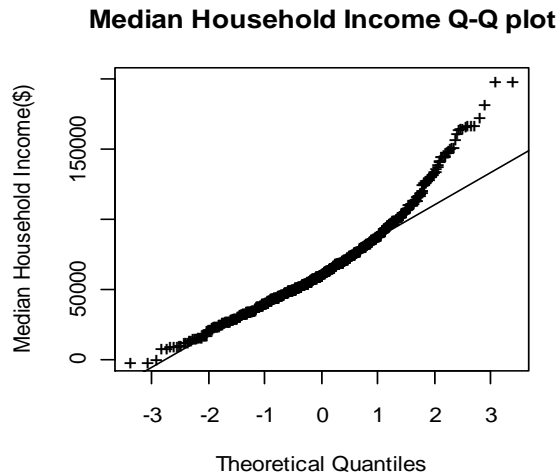
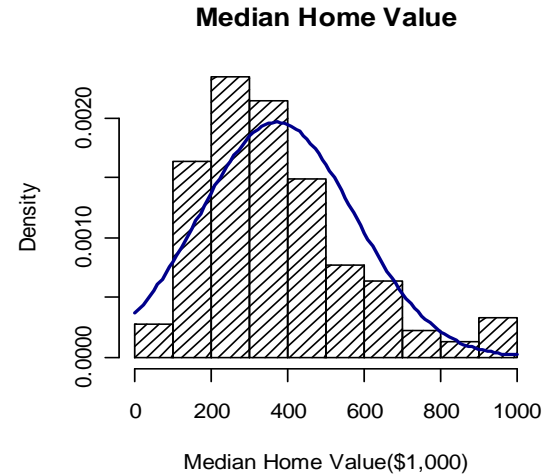
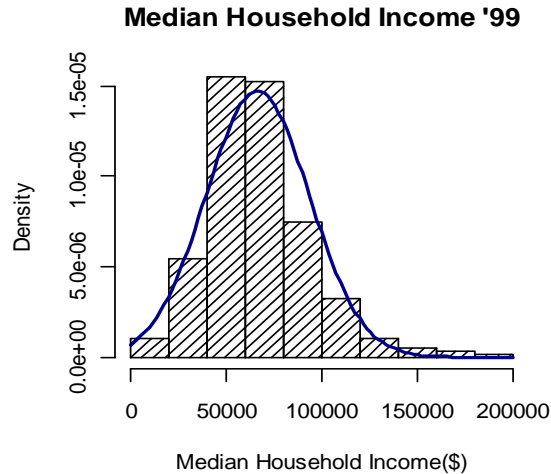


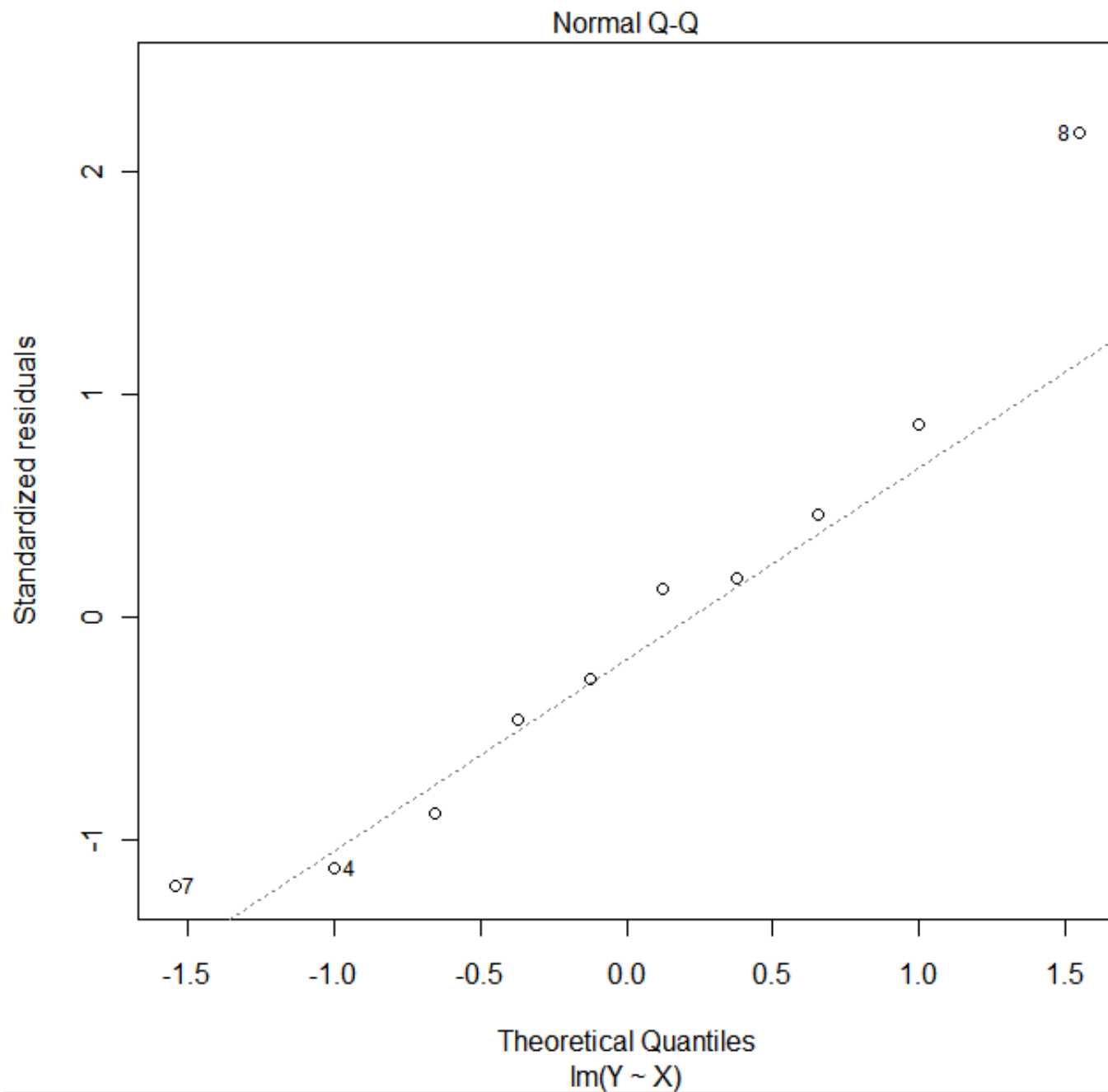
# Residual Plot

Dependent Variable: Wage in \$ per hour



# Normal Quantile-Quantile Plot





# Beta Weight (Coefficient)

- Regression coefficients when variables are standardized

$$\hat{Z}_Y = a + b_1^* Z_{X1} + b_2^* Z_{X2}$$

$b_1^*$  &  $b_2^*$  are beta weights (sometimes also notated  $\beta_1$  &  $\beta_2$ ). They reflect the relative strength of independent variables ( $X_1$  &  $X_2$ ) in predicting the dependent variable ( $Y$ ). If  $b_1^*$  is 3 times larger (in absolute terms) than  $b_2^*$ , then can say  $X_1$  has 3 times the explanatory power of  $X_2$ .

- Can also compute as:

$$b_1^* = b_1(S_{X1}/S_Y) \text{ for } \hat{Y} = b_0 + b_1X_1 + b_2X_2$$

Table 6. Regression model predicting vehicle miles of travel, with and without neighborhood walkability index (N = 5,710).

Independent variables	Unstandardized coefficients		Standardized coefficients	t	Sig.	Partial corr.	Variance explained (%)
	B	SE	Beta				
Constant	.988	.029		33.621	.000		
Gender	-.043	.011	-.050	-3.985	.000	-.050	0.25
Education	.057	.003	.253	19.787	.000	.248	6.13
Household income	.018	.003	.072	5.327	.000	.067	0.44
Vehicles per household	.022	.006	.054	3.906	.000	.049	0.24
Miles to nearest bus stop	.029	.009	.045	3.164	.002	.040	
Walkability index	-.019	.002	-.157	-10.740	.000	-.134	1.81

Source: Lawrence D. Frank, James F. Sallis, Terry L. Conway, James E. Chapman, Brian E. Saelens & William Bachman (2006): Many Pathways from Land Use to Health: Associations between Neighborhood Walkability and Active Transportation, Body Mass Index, and Air Quality, Journal of the American Planning Association, 72:1, 75-87