GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 3: LINEAR REGRESSION

AGENDA

I. REVIEW
II. INTRO TO REGRESSION PROBLEMS
III. BUILDING REGRESSION INTUITION
IV. LAB: PRACTICE

I. SOME REVIEW...

SUPERVISED LEARNING

- Outcome measurement ***Y***, (also called dependent variable, response, target)

SUPERVISED LEARNING

- Outcome measurement ***Y***, (also called dependent variable, response, target)
- Vector of ***p*** predictor measurements ***X*** (also called inputs, regressors, covariates, features, independent variables)

SUPERVISED LEARNING

- Outcome measurement ***Y***, (also called dependent variable, response, target)
- Vector of ***p*** predictor measurements ***X*** (also called inputs, regressors, covariates, features, independent variables)
- In ***regression***, *Y* is quantitative (e.g. price, temperature)

II. INTRO TO REGRESSION PROBLEMS

LINEAR REGRESSION ...

- How does sales volume change with changes in price? How is this affected by changes in weather?
- Is there a relationship between the amount of a drug absorbed and body weight of a patient?
- Can we explain the effect of education on income?
- How does the energy released by an earthquake vary with the depth of its epicenter?

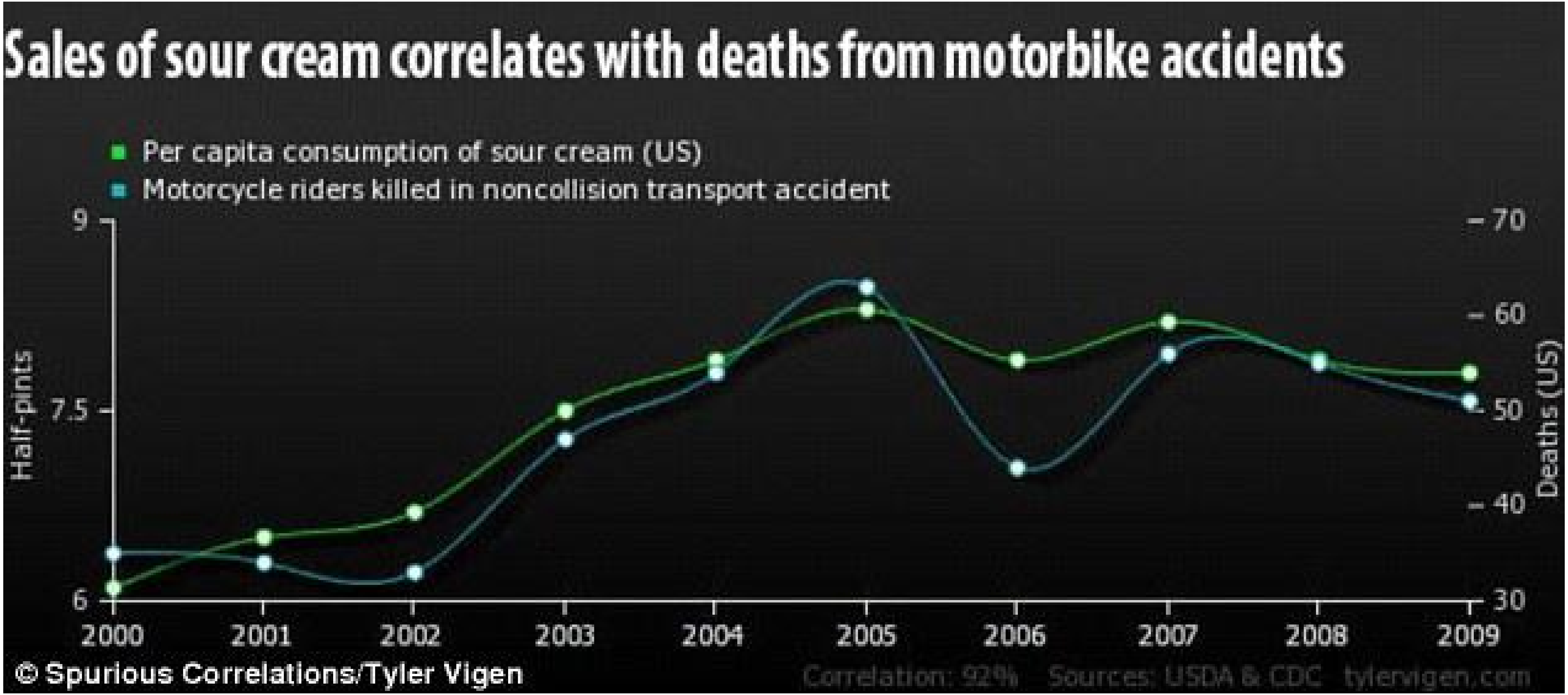
LINEAR REGRESSION ...

- is used to predict future outcomes and understand relationships
- is a simple approach to supervised learning
- may seem overly simplistic, but is extremely useful both conceptually and practically

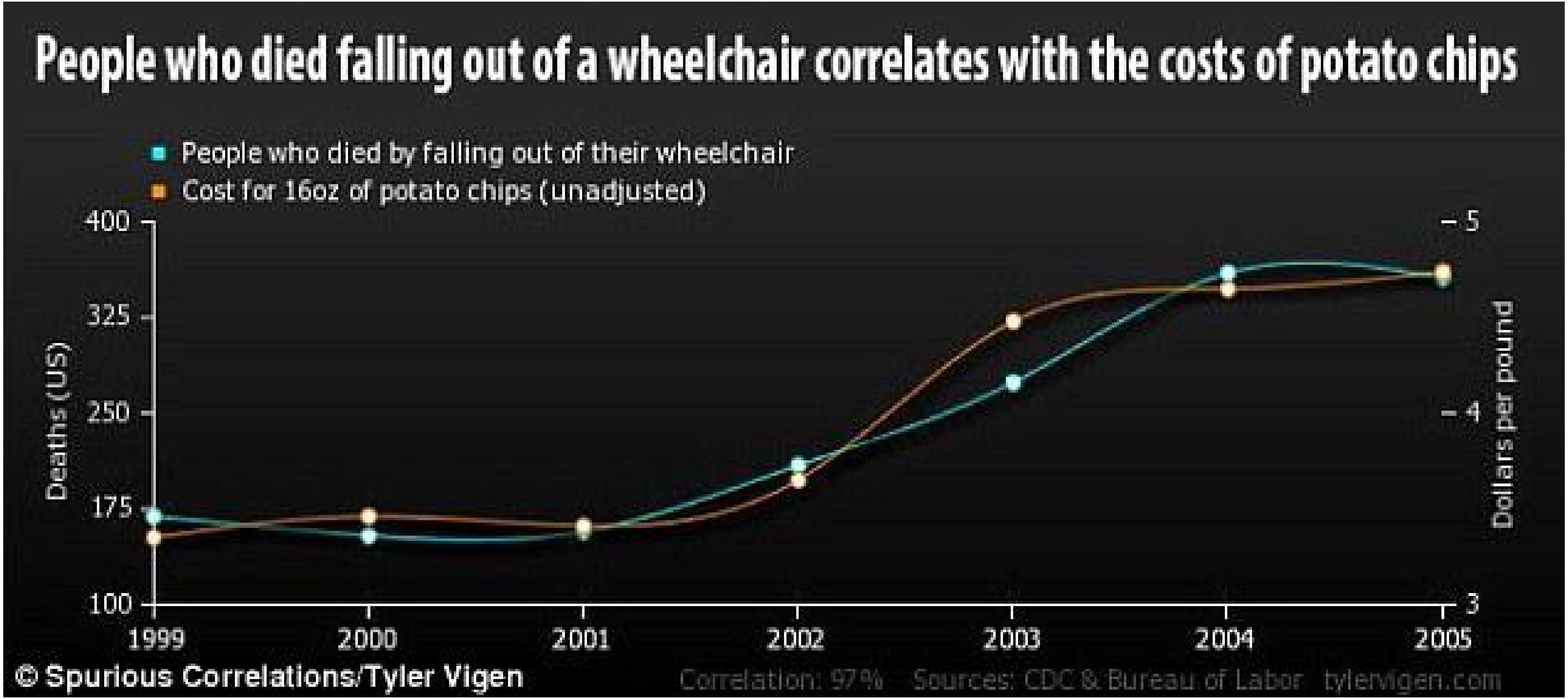
Linear regression ...

- the dependent variable is a ***continuous*** variable
- the independent variable(s) can take any form - continuous or discrete
- does not establish a cause-and-effect relationship-- just that there is a relationship

LINEAR REGRESSION …



LINEAR REGRESSION ...



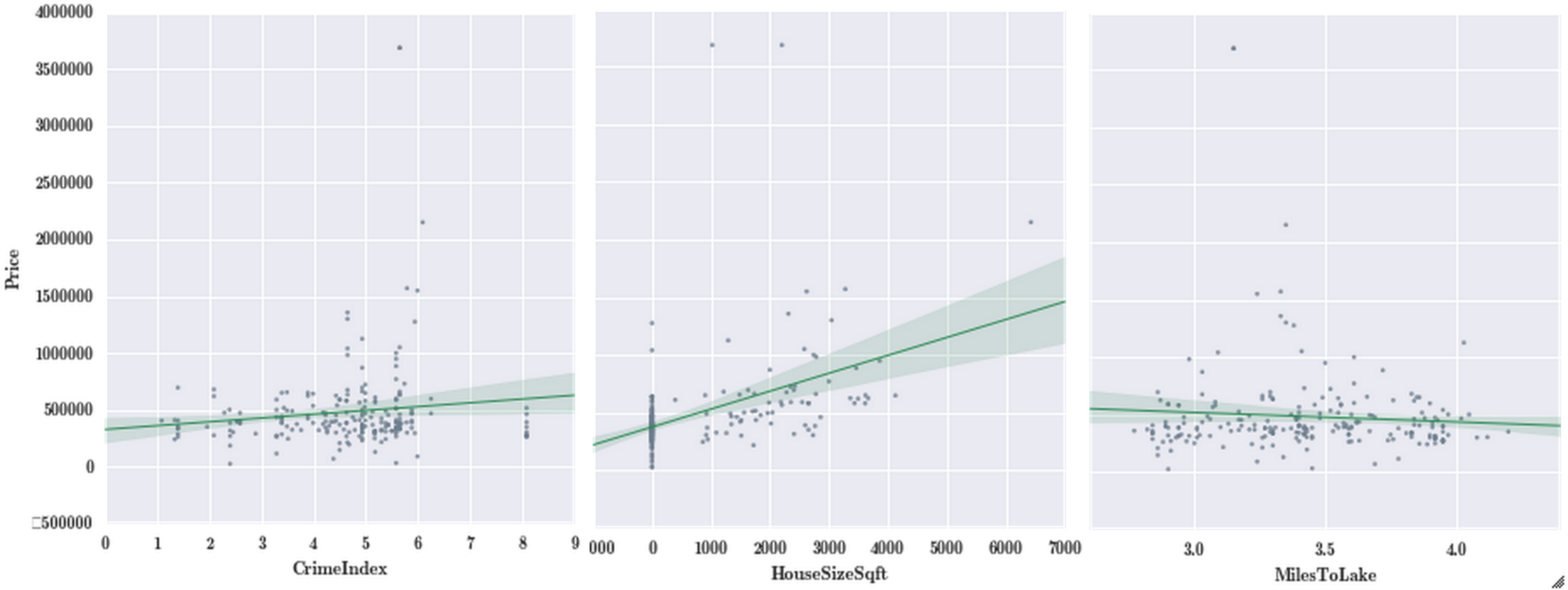
LINEAR REGRESSION ASSUMPTIONS

- The relationship between the variables is ***linear***.
- The data is ***homoskedastic***, meaning the variance in the ***residuals*** (the difference in the real and predicted values) is more or less constant
- The ***residuals*** are independent (distributed randomly and not influenced by the residuals in previous observations). If not, they are ***autocorrelated***

Independent and identically distributed?

- A sequence of outcomes of spins of a roulette wheel
- A sequence of daily weather conditions
- A sequence of fair or loaded dice rolls
- A sequence of daily stock prices
- A sequence of fair or loaded coin flips

CONSIDER THE FOLLOWING DATASET:



QUESTIONS WE MIGHT ASK ABOUT THIS DATA

- Is there a relationship between *House Size* and *Price*?
- How strong is the relationship between *Crime Index* and *Price*?
- Which features contribute most to *Price*?
- How accurately can we predict future Prices?
- Is the relationship linear?
- Is there any synergy among the different features?

LINEAR REGRESSION MODEL

Q: What is a regression model?

LINEAR REGRESSION MODEL

Q: What is a regression model?

A: A functional relationship between input and response variables

LINEAR REGRESSION MODEL

Q: What is a regression model?
A: A functional relationship between input and response variables

The **simple linear regression** model captures a linear relationship between a single input variable X and a response variable Y:

LINEAR REGRESSION MODEL

Q: What is a regression model?
A: A functional relationship between input and response variables

The **simple linear regression** model captures a linear relationship between a single input variable X and a response variable Y:

$$y = \alpha + \beta x$$

LINEAR REGRESSION MODEL

Q: What do the terms in this model mean?

$$y = \alpha + \beta x$$

LINEAR REGRESSION MODEL

Q: What do the terms in this model mean?

$$y = \alpha + \beta x$$

A:

y = response variable (target, dependent variable)

LINEAR REGRESSION MODEL

Q: What do the terms in this model mean?

$$y = \alpha + \beta x$$

A:

y = response variable (target, dependent variable)
x = input variable (feature, independent variable)

LINEAR REGRESSION MODEL

Q: What do the terms in this model mean?

$$y = \alpha + \beta x$$

A:

y = response variable (target, dependent variable)
x = input variable (feature, independent variable)
α = constant bias term, y-intercept

LINEAR REGRESSION MODEL

Q: What do the terms in this model mean?

$$y = \alpha + \beta x$$

A:

y = response variable (target, dependent variable)
x = input variable (feature, independent variable)
α = constant bias term, y-intercept
β = regression coefficient (model parameter)

ORDINARY LEAST SQUARES (OLS) METHOD

$$SS_{res} = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

- A cost function is used to measure the error of model

ORDINARY LEAST SQUARES (OLS) METHOD

$$SS_{res} = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

- A cost function is used to measure the goodness-of-fit of model
- The values of the model parameters that minimize the cost function produce the best model.

ORDINARY LEAST SQUARES (OLS) METHOD

$$SS_{res} = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

- A cost function is used to measure the goodness-of-fit of model
- The values of the model parameters that minimize the cost function produce the best model.
- The **residual sum of squares** cost function sums the squares of the **residuals**, or training errors.

ORDINARY LEAST SQUARES (OLS) METHOD

SOLVING FOR BETA

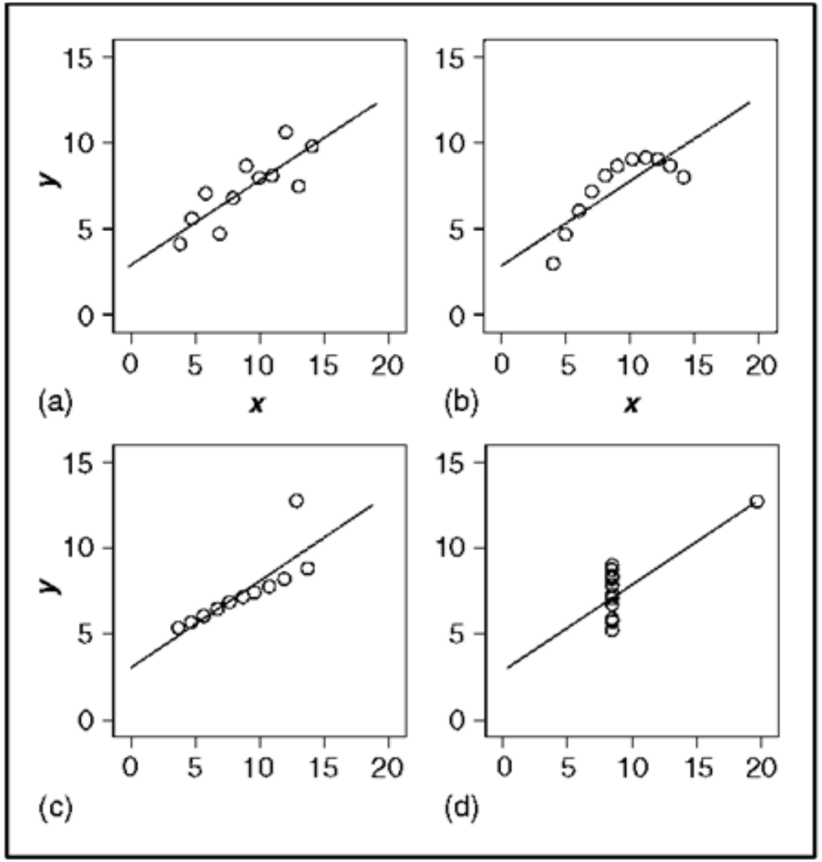
For simple linear regression, the slope of the regression line (beta) is equal to the corrected correlation between the explanatory variable and the response variable.

$$\beta = \frac{cov(x, y)}{var(x)}$$

SOLVING FOR ALPHA

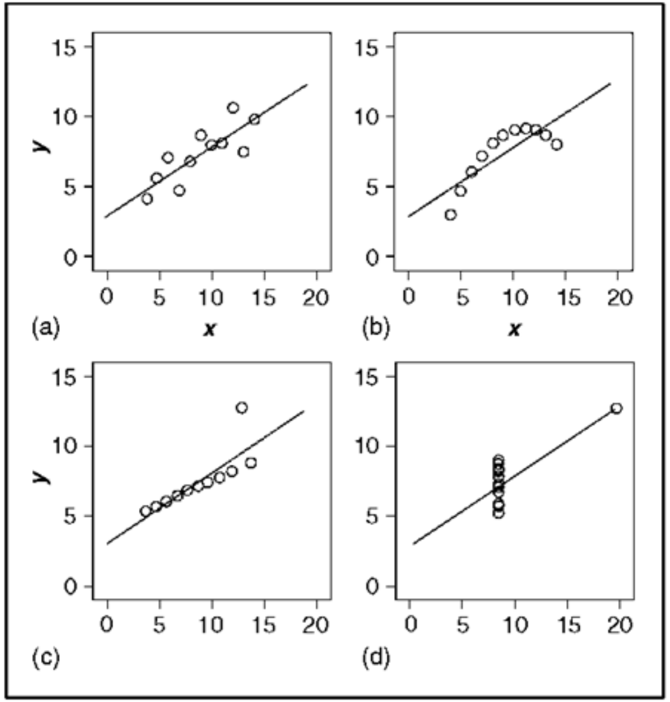
$$\alpha = \bar{y} - \beta\bar{x}$$

LINEAR REGRESSION GOTCHAS



LINEAR REGRESSION

same least-squares regression, regression coefficients, standard errors, correlation between variables, and standard error!!



AN EXAMPLE

III. BUILDING REGRESSION INTUITION

IV. Lab: Practice