GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

THE LINEAR REGRESSION

I. INTRODUCTION TO REGRESSION DATA PROBLEMS
II. HOW REGRESSIONS WORK
III. DETERMINING COST

EXERCISES:
IV. IMPLEMENTING THE LINEAR MODEL

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)

I. LINEAR REGRESSION

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?

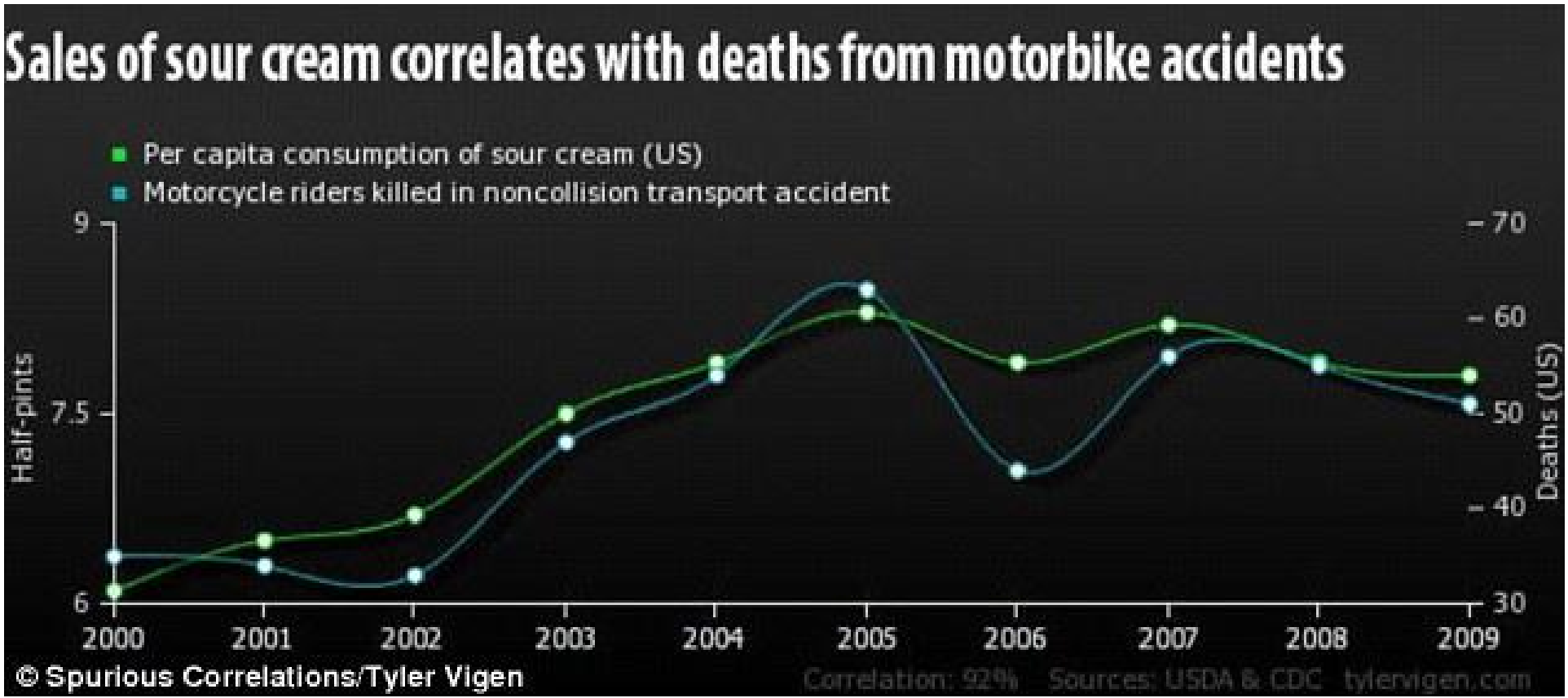
LINEAR REGRESSION ...

- 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

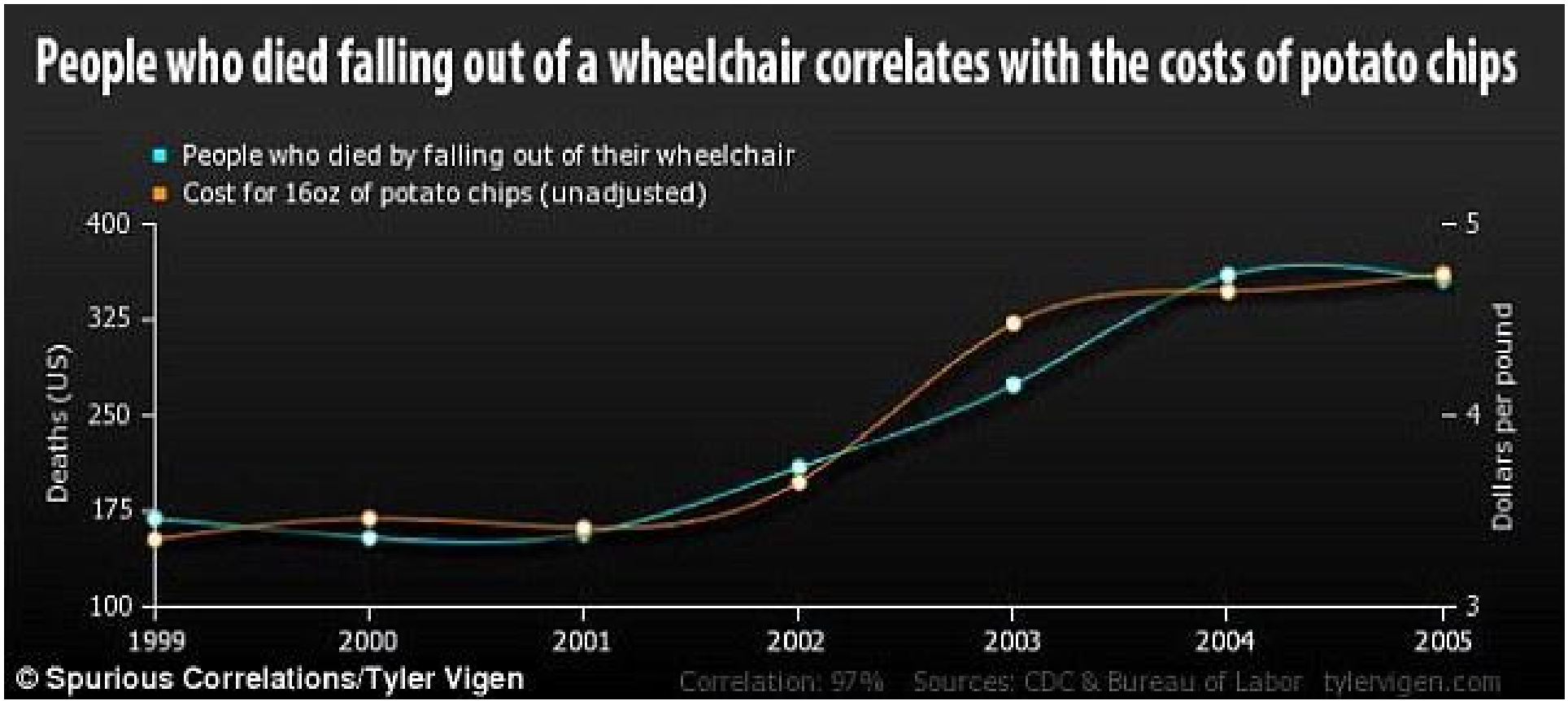
LINEAR REGRESSION ...

- 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 ...



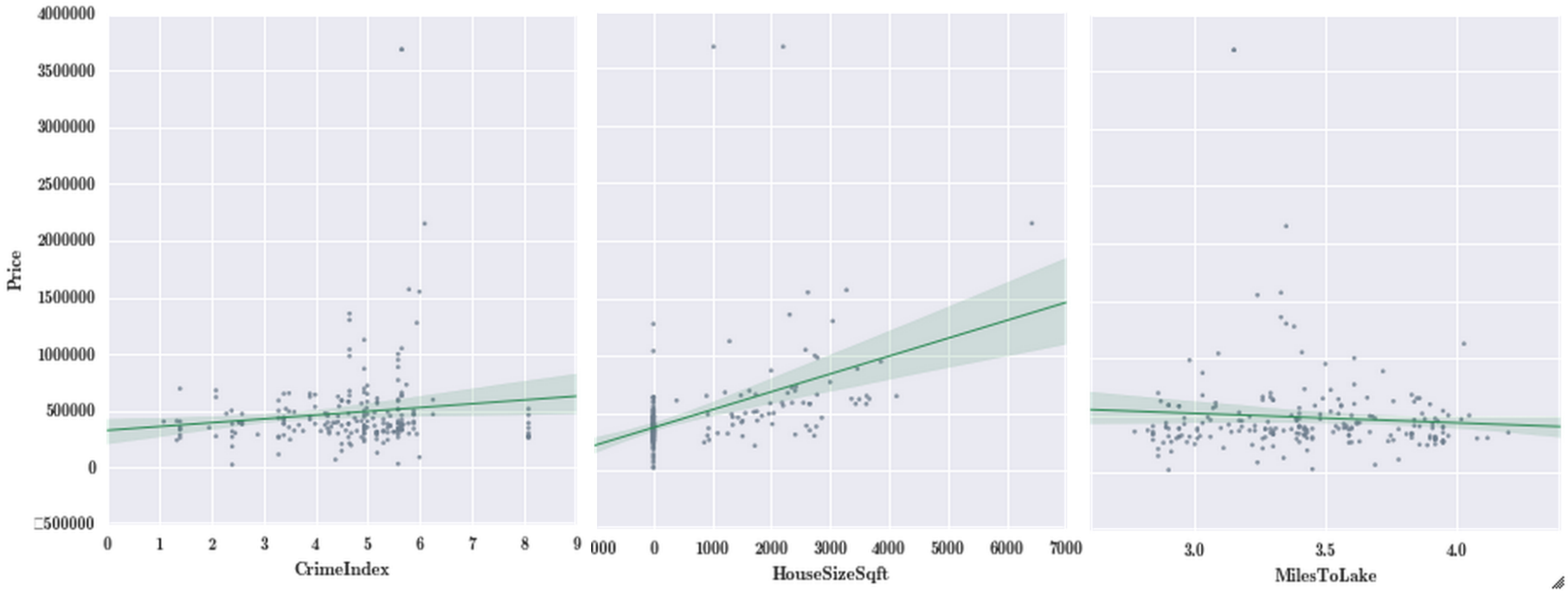
LINEAR REGRESSION ASSUMPTIONS

- 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

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

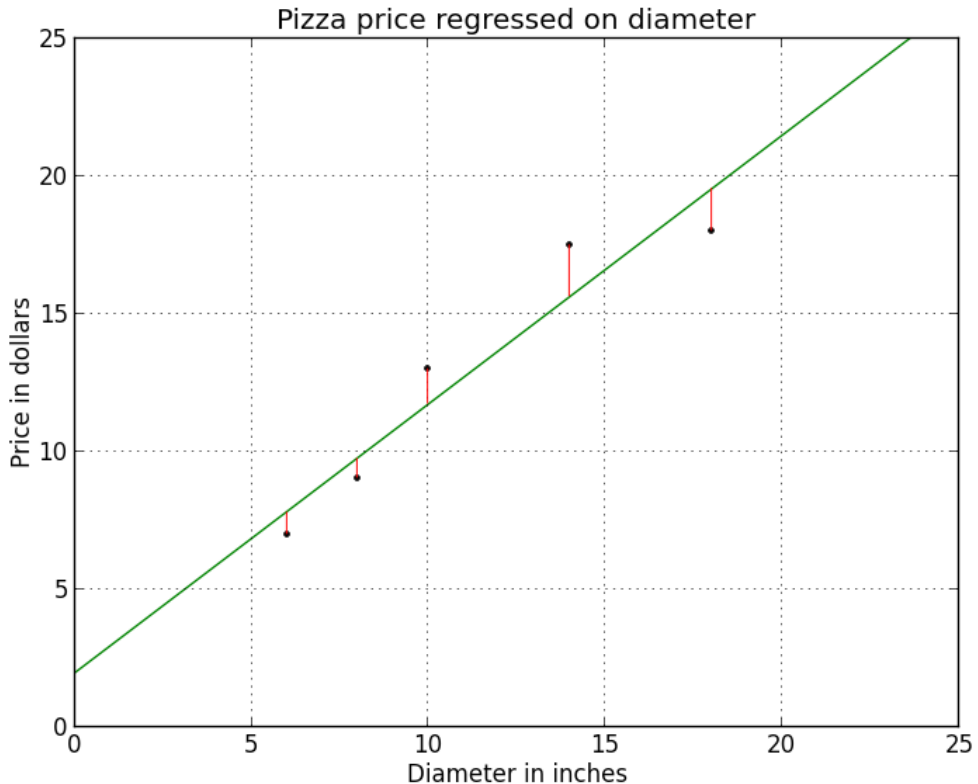
- A cost function is used to measure the goodness-of-fit of model
- The values of the model parameters that minimize the cost function produce the best model.

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

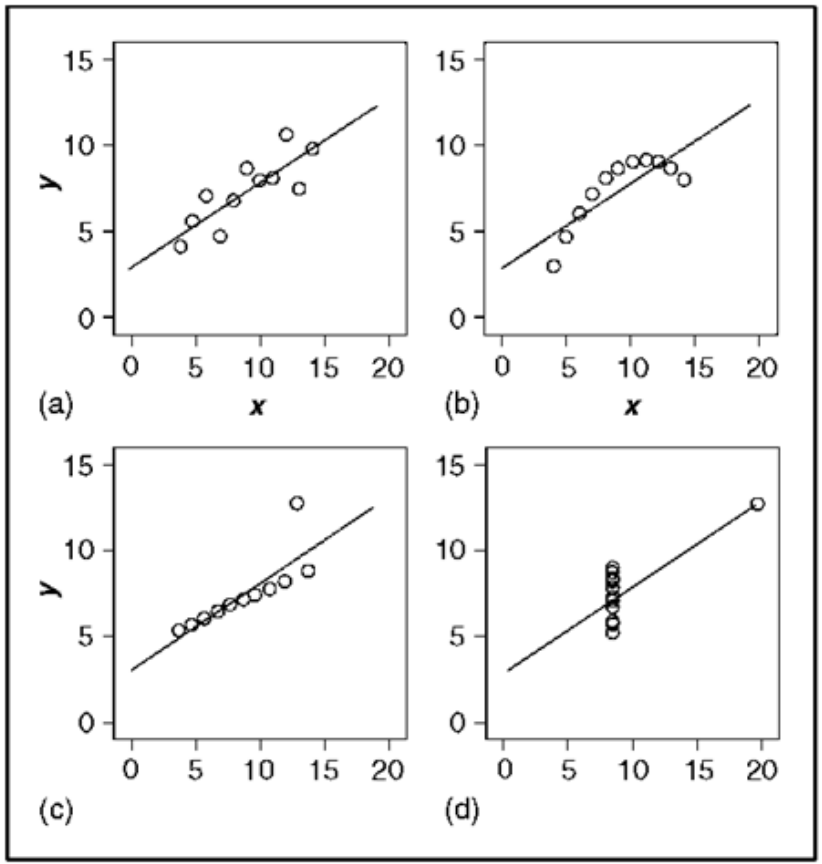
For simple linear regression, the slope of the regression line (beta) is equal to the corrected correlation between the explanatory variable and the response variable.

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

SOLVING FOR ALPHA

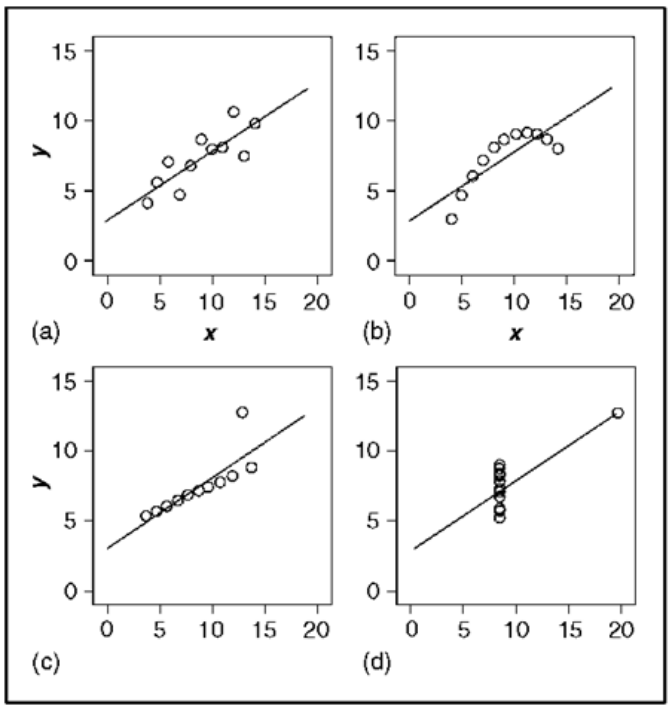
$$\alpha = \bar{y} - \beta\bar{x}$$

LINEAR REGRESSION GOTCHAS



LINEAR REGRESSION

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



INTRO TO DATA SCIENCE

II: POLYNOMIAL REGRESSION

Consider the following **polynomial regression** model:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Consider the following **polynomial regression** model:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Q: This represents a nonlinear relationship. Is it still a linear model?

Consider the following **polynomial regression** model:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Q: This represents a nonlinear relationship. Is it still a linear model?

A: Yes, because it's linear in the β's!

Consider the following **polynomial regression** model:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Q: This represents a nonlinear relationship. Is it still a linear model?

A: Yes, because it's linear in the β's!

"Although polynomial regression fits a *nonlinear* model to the data, as a statistical estimation problem it is *linear*, in the sense that the regression function E(y|x) is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression." -- Wikipedia

Polynomial regression allows us to fit very complex curves to data.

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon$$

Polynomial regression allows us to fit very complex curves to data.

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon$$

But there is one problem with the model we've written down so far.

Polynomial regression allows us to fit very complex curves to data.

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon$$

But there is one problem with the model we've written down so far.

Q: Does anyone know what it is?

Polynomial regression allows us to fit very complex curves to data.

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon$$

But there is one problem with the model we've written down so far.

Q: Does anyone know what it is?

A: This model violates one of the assumptions of linear regression!

This model displays **multicollinearity**, which means the predictor variables are highly correlated with each other.

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon$$

Multicollinearity causes the linear regression model to break down, because it can't tell the predictor variables apart.

III: COST OF LINEAR REGRESSIONS

Q: How do measure error in a linear regression model?

Q: How do measure error in a linear regression model?
A: In theory, minimize the sum of the squared residuals (RSS, or SSE).

Q: How do measure error in a linear regression model?
A: In theory, minimize the sum of the squared residuals (RSS, or SSE).

In practice, any respectable software can do this for you.

Q: How do measure error in a linear regression model?
A: In theory, minimize the sum of the squared residuals (RSS, or SSE).

In practice, any respectable software can do this for you.

In python, we can find this with some quick code.

Q: How do measure error in a linear regression model?
A: In theory, minimize the sum of the squared residuals (RSS, or SSE).

In python, we can find this with some quick code:

$$\text{mean}((\text{prediction} - \text{actual})^2)$$

Q: How do measure goodness of fit?
A: In theory, we want to maximize R^2 (as close to one as possible).

Q: How do measure goodness of fit?
A: In theory, we want to maximize R^2 (as close to one as possible).

Sklearn already calculates this for us, as do any other stats packages and programs.

Q: How do measure goodness of fit?
A: In theory, we want to maximize R^2 (as close to one as possible).

Sklearn already calculates this for us, as do any other stats packages and programs.

If you want to get serious into regression, learn more about the coefficient of determination.

HYPOTHESIS TESTING: P-VALUE

Coin example. Friend gets 10 rounds in a row heads.

- Our null hypothesis is that the coin is fair.
- *p*-value: ~0.2% chance of the effect we observed happening
- So, we reject the null hypothesis of a fair coin and claim it is unfair.

P-value: How likely is the effect observed in the sample data, given that the null hypothesis is true?

HYPOTHESIS TESTING: P-VALUE

Vaccine example. Are vaccines effective?

- *p*-value: 0.04 = 4%.
- If the vaccine has no effect, you'd see this difference in no more than 4% of studies due to random sampling error.

- P-value: How likely is your data, assuming a true null hypothesis?
- **P-values are not the error rate.**

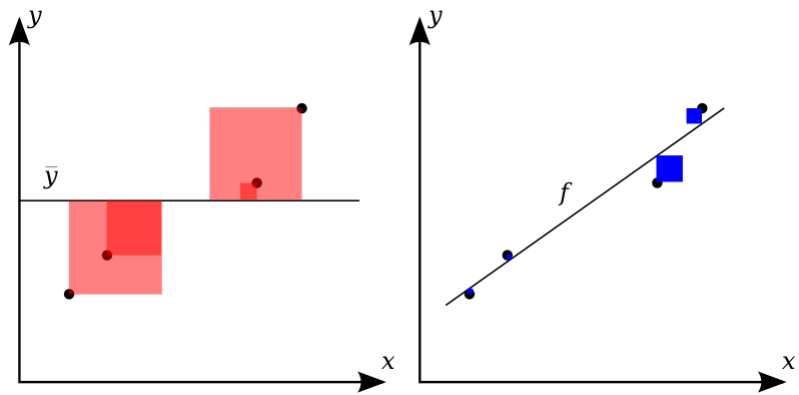
See: http://blog.minitab.com/blog/adventures-in-statistics/how-to-correctly-interpret-p-values

R-SQUARED VALUE

Coefficient of determination: R^2

- In linear regression, R^2 is the square of the sample correlation coefficient. Interpreted as the proportion of response variation explained by the regressors in the model. The better the linear regression fits the data as opposed to the average, the closer to 1 it is.
- Related to the unexplained variance:

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$



- Adjusted R^2: Adjusts for # explanatory terms with respect to # datapoints

F-STATISTIC (FOR MODEL AND/OR COEFF)

- “Fixation index”
- F-statistic/P(F-statistic): Tests the statistical significance of the model. Does the model fit the data better than the mean?
- The p-value null hypothesis is whether $R^2 = 0$ should be rejected. Compares how this data fits vs the same dataset with less information.

- F-statistic can mean the model is statistically significant, even though it may only explain a low variability in the response (R^2)

T-STATISTIC (PER COEFFICIENT)

- Tests the null hypothesis that the term is not significant (i.e. it is equal to zero)

- If the R^2-value for the linear regression is significant but the p-value of the coefficients are not significant, there may be multicollinearity among the predictor variables.