NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Week 10: Lecture 2

Linear Regression – Relating Attributes of Data

NYU Center for Data Science

How can we use a known quantity to predict an unknown quantity?

DS-UA 112
Introduction to Data Science

Week 10: Lecture 2

Linear Regression – Relating Attributes of Data

Adapted from Nolan, Speed, Gonzalez, Lau

Announcements

- Please check Week 10 agenda on NYU Classes
  - Lab 9
    - Due on Friday April 3 at 12PM
  - Project 1
    - Due on Monday April 6 at 12PM
  - Survey

https://nyu.qualtrics.com/jfe/form/SV_3DCWUa4yc08L0wt

Review

Question 1

We use probability in modelling to characterize randomness and quantify _____ due to sampling.

- observations
- uncertainty
- probability
- error

Question 2

Which of the following are true of random variables taking discrete values?

- They can be added to other random variables
- They can be multiplied by other random variables
- They are denoted by capital letters
- They have an associated expectation and variance

Review

Question 3

Which of the following are True?

- E[X + Y] = E[X] + E[Y] for all random variables X and Y.
- E[X +Y] = E[X]+E[Y] for independent random variables X and Y.
- Var[X + Y] = Var[X] + Var[Y] for all random variables X and Y.
- Var[X + Y] = Var[X] + Var[Y] for independent random variables X and Y.

Review

Question 4

An estimator is a function

- of a statistic that computes a parameter.
- of a distribution that computes a statistic.
- of a random variable that computes its expectation.
- of a sample that computes an estimate of a parameter.

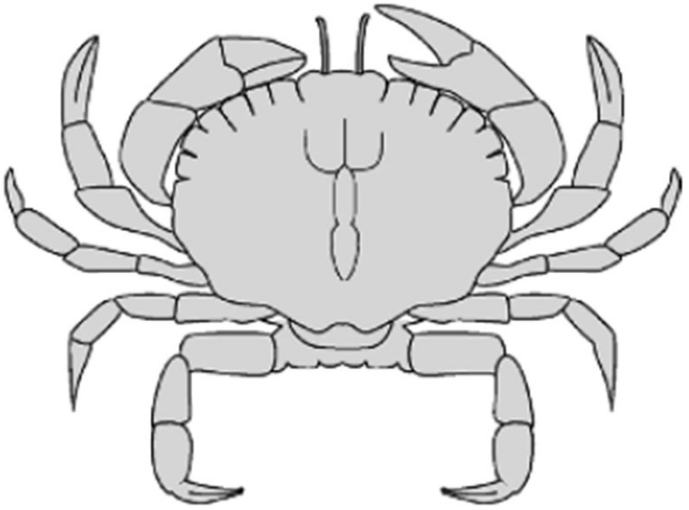
Question 5

Which of the following give the variance of a random variables?

- $(X - \theta)^2$
- $E[X - \theta]^2$
- $E[(X - \theta)^2]$
- $(E[X] - \theta)^2$

Agenda

- Correlation
  - Capturing liner relationship
- Linear Regression
  - Prediction unknown data from known data
- Example



Dungeness Crab (*Cancer magister*).

Correlation

- Suppose we want to study two quantitative variables about a population.
- Using simple random sampling we can generate datasets

$$\{y_1, \ldots, y_n\}$$

$$\{x_1, \ldots, x_n\}$$

- We could think of each sample as a possible value of random variables X and Y
- With information about X, could we guess information about Y?

- Estimation means studying the joint distribution

P(X=x and Y=y)

- Prediction means studying the conditional distribution

P(Y=y | X=x)

Correlation

- If it appears the X and Y differ by a transformation that involves scaling

$$X \longrightarrow c\,X$$

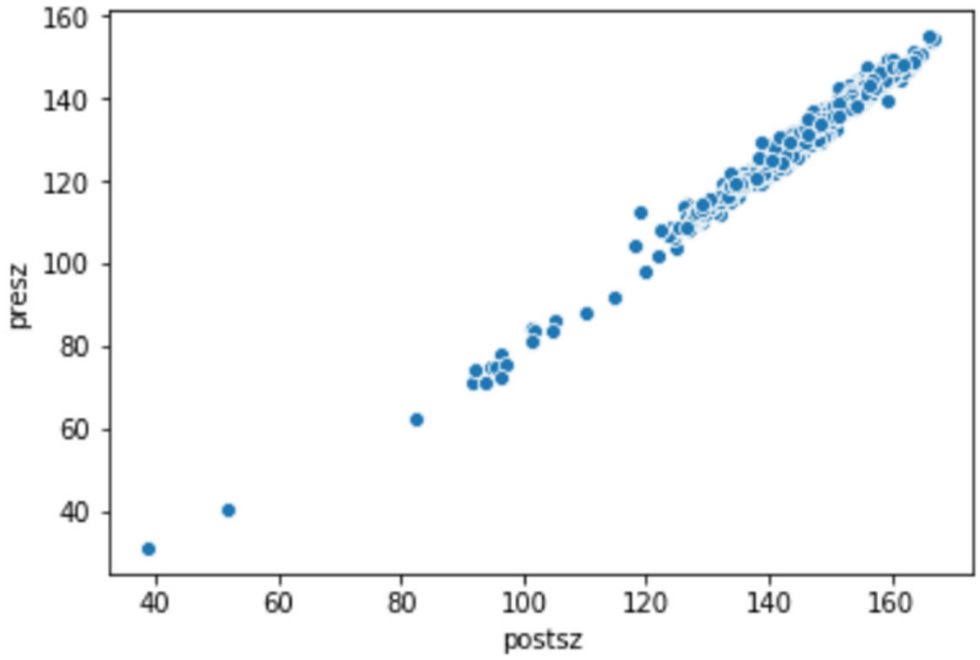
and shifting

$$c\,X \longrightarrow c\,X + d$$

then maybe we can find a linear relationship between the random variables.

- Correlation measures the linear relationship between random variables

- Define the covariance of X and Y to be E[(X –E[X])(Y – E[Y])]
- Note that for X = Y we have the definition of variance



Correlation

- We need to know the probability distribution of the population to compute expectation in the definition of covariance
- Instead we use the data generated in the sample to guess the correlation.
- For the datasets we define the correlation to be

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

- Note that we denote the mean with

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad \overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

- Note that the standard deviations of the datasets are

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

- The remaining expression in the numerator is called the covariance of the datasets

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2$$

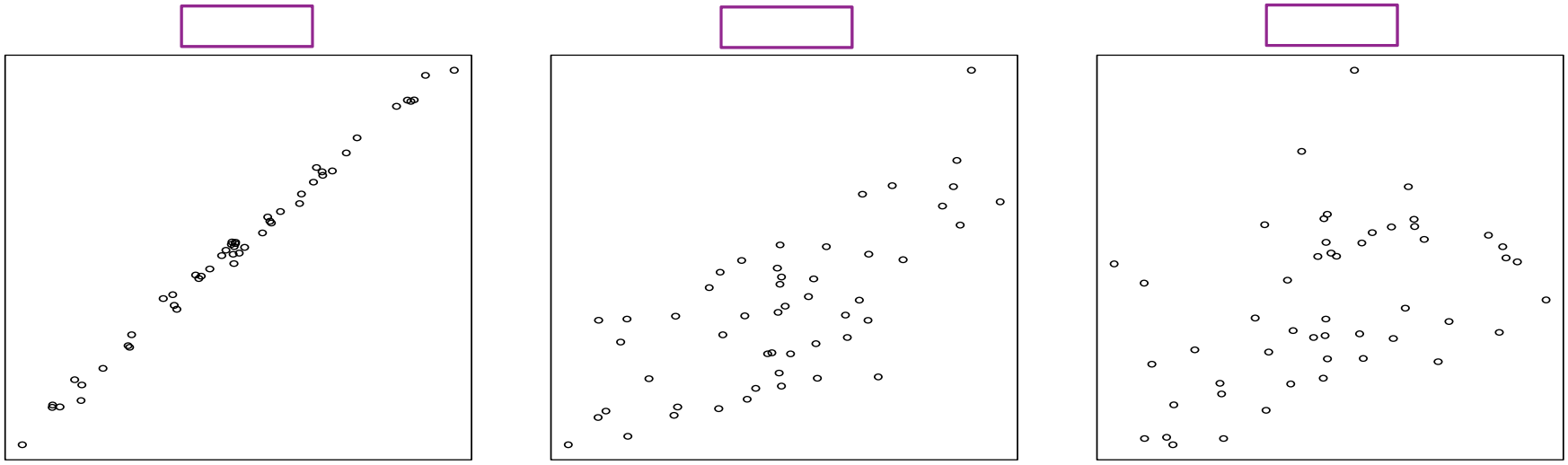
Exercise

- We have three scatter-plots. Suppose the correlations for each pairs of datasets is a number in

  0.95, 0.75, 0.50, 0.30, 0.10

- Label the three charts

- Note that we scaled the datasets to have standard deviation 1. Without scaling the datasets to have standard deviation 1, we might miss the linear trend between the datasets because the scales might be different

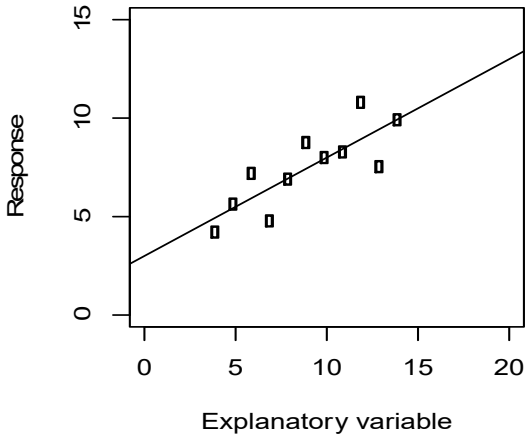


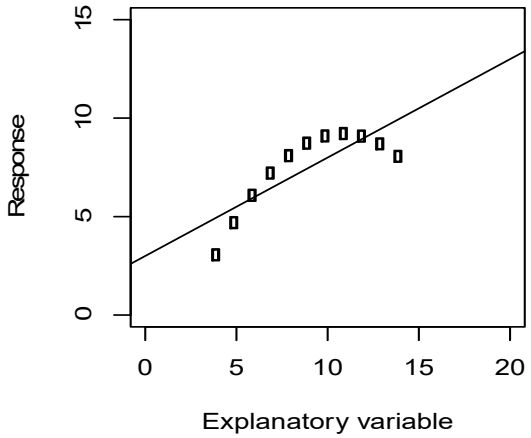
Correlation

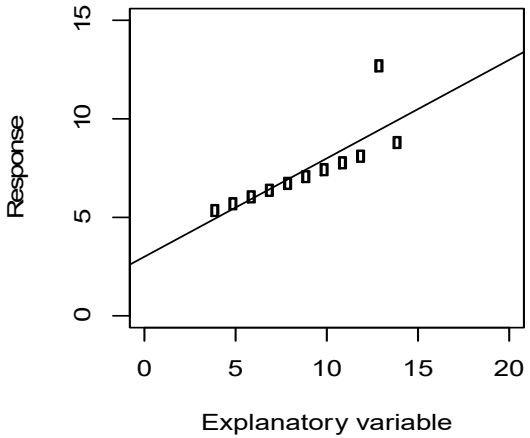
- If we can associate units to the random variables X and Y like millimeters then we find that the units from the numerator cancel the units from the denominator.
- The value of the correlation lies between -1 and 1.
  - Positive correlation means values near 1.
  - Negative correlation means values near -1.
  - Uncorrelated means values near 0
- While correlation suggests a relationship between X and Y, we cannot assert that positive / negative correlation indicates X causes Y.
- Remember that we obtained the datasets through a simple random sample meaning we had an observational study.
- If we want to assess causation, then we would need an experimental study where we could control the presence / absence of features of X to measure the response in Y.

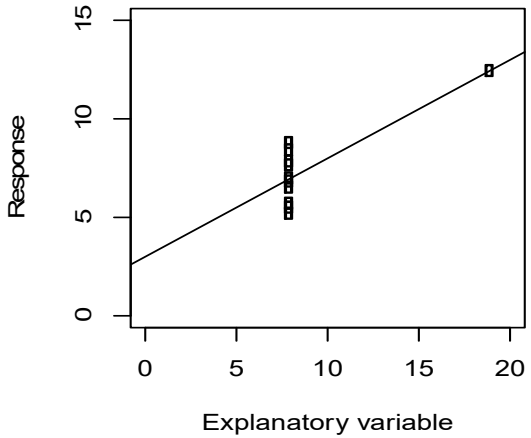
Correlation

- Remember the Anscombe Quartet from the lectures on visualization. These datasets have the same correlation
- Correlation cannot capture nonlinear trends. For example, $Y = X^2$ would mean a causal relationship between X and Y. However the correlation may not be 1 / -1









Exercise

- Suppose we have a table consisting of student grades on an assignment. The columns indicate the
  - score on assignment
  - points deducted by graders
- We know that the datasets have a causal relationship and a linear relationship.
- How can we compute the correlation?

score	Points lost
25	0
20	5
22	3
15	10
25	0

Exercise

- Note the linear relationship

  score = 25 – points lost

- Using the expression we can relate the means

$$\bar{y} = \frac{1}{n}\sum_i (25 - x_i) = 25 - \bar{x}$$

and the variances

$$Var(y) = \frac{1}{n}\sum_i [25 - x_i - (25 - \bar{x})]^2 = Var(x)$$

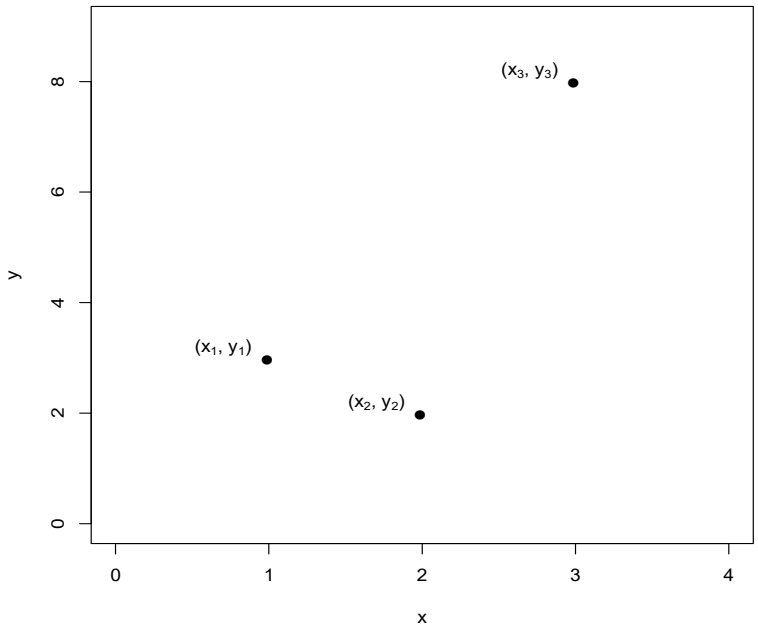
$$\begin{aligned} r &= \frac{1}{n}\sum_i \frac{x_i - \bar{x}}{SD(x)}\frac{y_i - \bar{y}}{SD(y)} \\ &= \frac{1}{n}\sum_i \frac{x_i - \bar{x}}{SD(x)}\frac{\bar{x} - x_i}{SD(x)} \qquad = -1 \end{aligned}$$

Fitting Lines to Data

- If the correlation is positive or negative, then we can try to capture the linear relationship between the datasets with a line.
- The formula for a line is

  $$y = a + b\,x$$

- Here a is the intercept. It shifts the line up or down.
- Here b is the slope. It mean relates the change in the vertical direction to the change in the horizontal direction

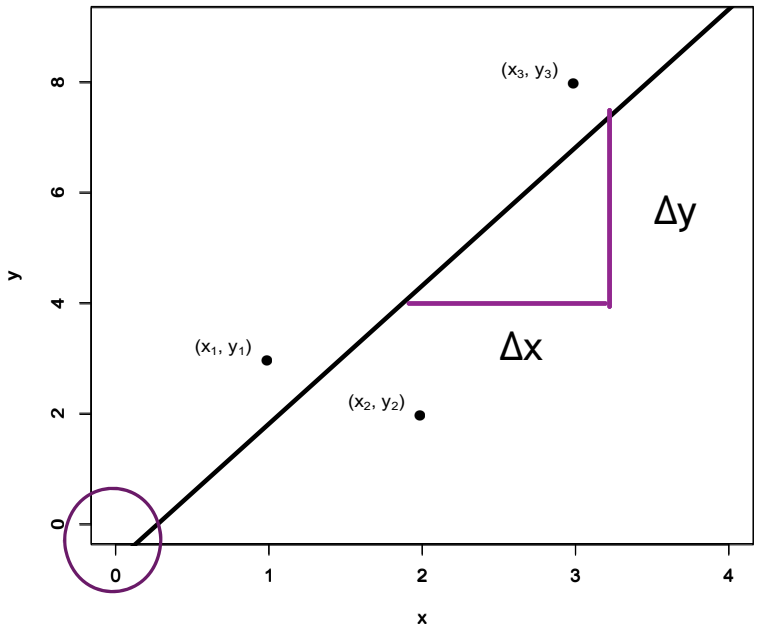


Fitting Lines to Data

- If the correlation is positive or negative, then we can try to capture the linear relationship between the datasets with a line.
- The formula for a line is

  $$y = a + b\,x$$

- Here a is the intercept. It shifts the line up or down.
- Here b is the slope. It mean relates the change in the vertical direction to the change in the horizontal direction $\Delta y / \Delta x$



Mean Square Error

- Unlike the constant model, we have two parameters in the linear model namely the slope and intercept.
- If we take the square loss function then the risk is

$$E[(Y - (a + b\,X))^2]$$

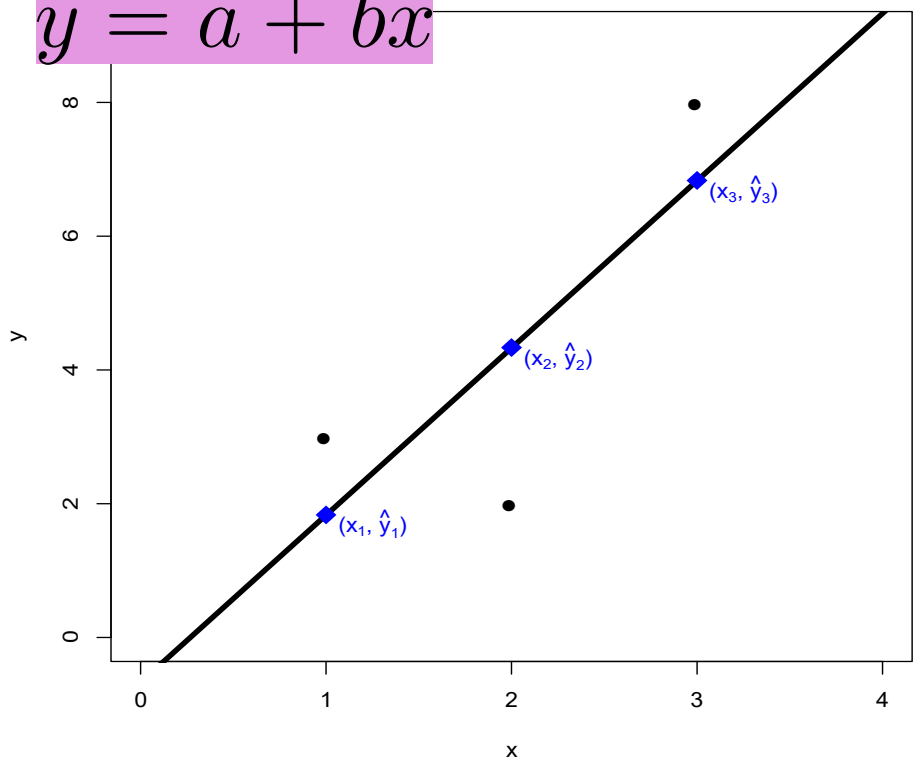
- We need to know the joint distribution of X and Y to compute the expectation. Instead we use the data to minimize the empirical risk

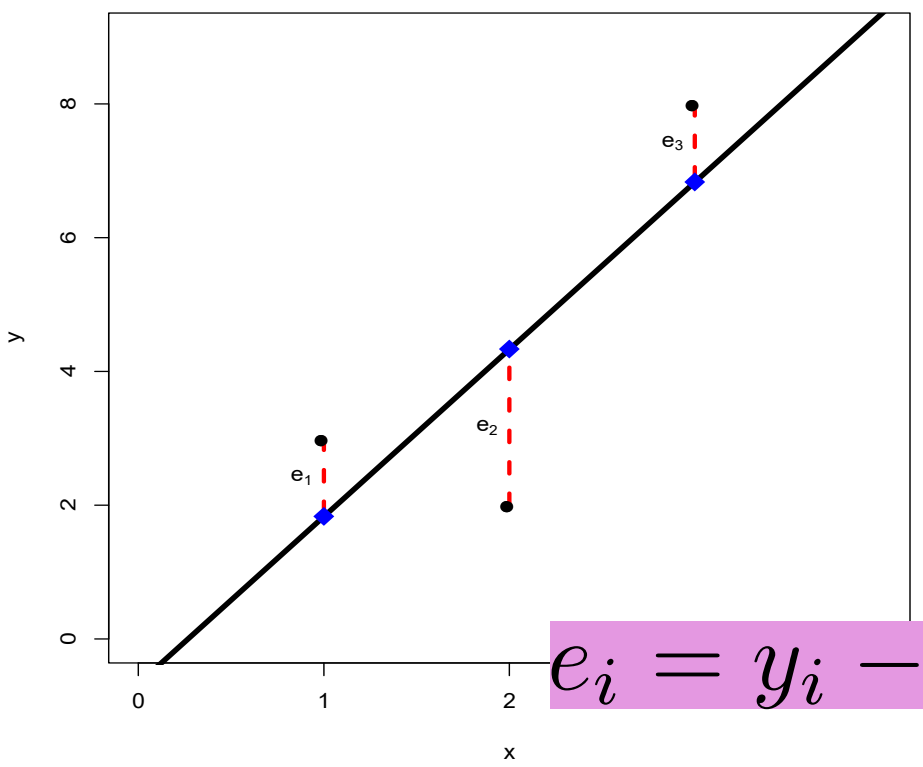
$$\min_{a,b} \frac{1}{n} \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

- Note that the loss function depends on the parameters a,b along with the data $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$
- Empirical means based on observations. Here we take the mean of the loss function over the observations in the datasets

Residuals

$$\hat{y} = \hat{a} + \hat{b}x$$

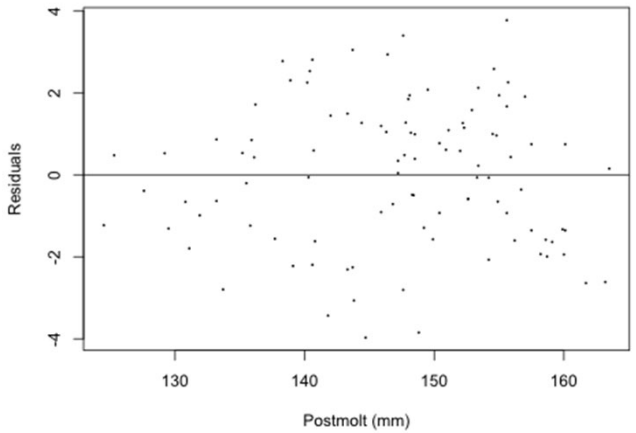




$$e_i = y_i - \hat{y}_i$$

Residuals

- For linear regression we should
  1. generate a scatter-plot showing the linear relationship among the points (x_i, y_i)
  2. determine slope and intercept that minimize the mean square error
  3. compute the residuals and generate another scatter-plot



- If the scatter-plot of the residuals contains any patterns, then we may need to
  - apply a transformation to feature variable or the response variable to capture a non-linear trend
  - use multiple feature variables

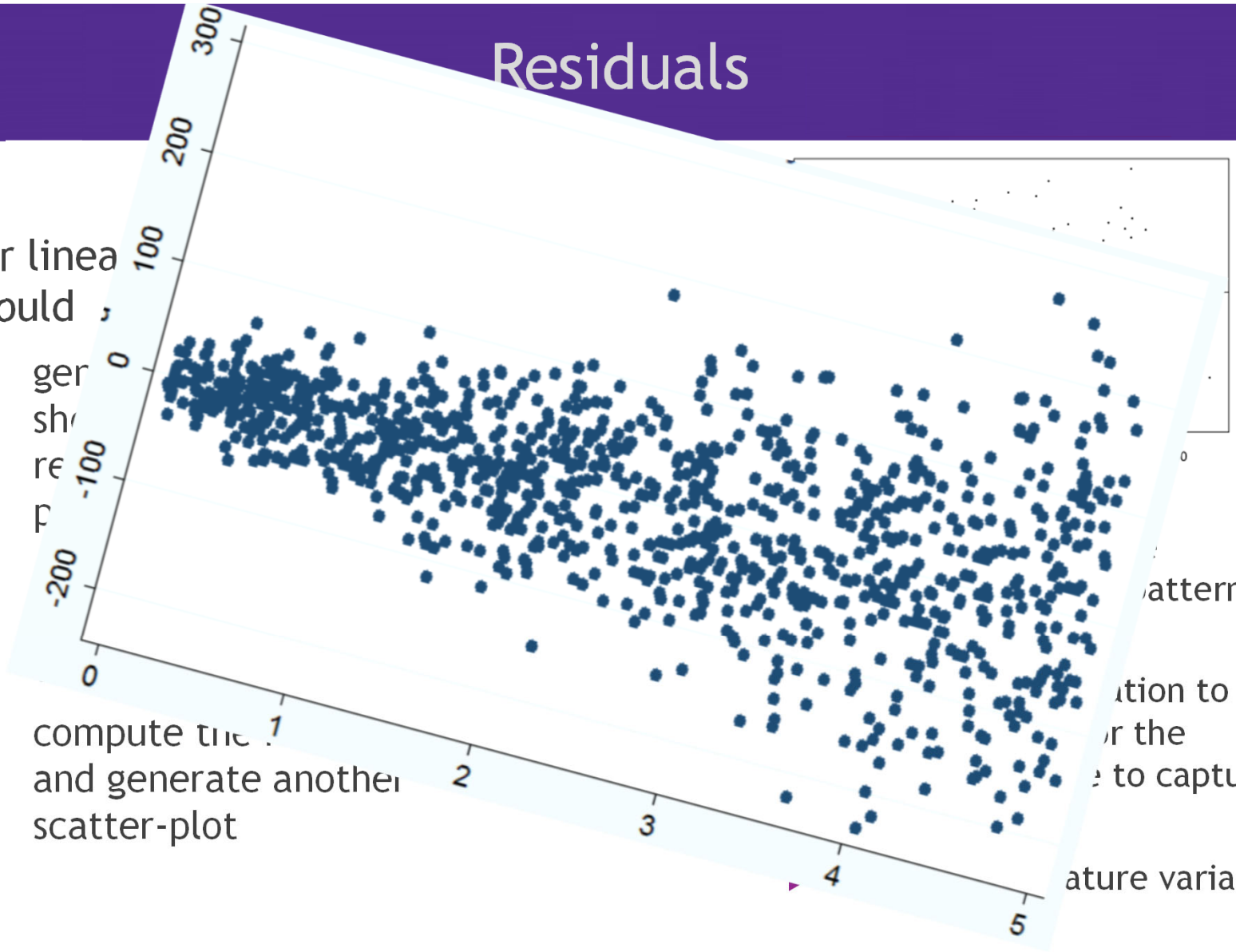
Residuals

- For linea
  should

1. ger
   sh
   re
   p
2. atterns,
3. compute the
   and generate another
   scatter-plot

tion to
r the
to capture a

ature variables



Minimizing Empirical Risk

- We can use derivative to find the parameters that minimize the mean square error

$$\min_{a,b} \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

- Note that we can omit the factor of $1/n$ because it does not affect the values of a,b that minimize the expression

- Derivative with respect to *a*

$$-2 \sum_i (y_i - a - bx_i)$$

- Derivative with respect to b

$$-2 \sum_i (y_i - a - bx_i) x_i$$

- Set to 0 and solve for *a* and *b*. Note that the equation for a shows the summation of residuals is 0

Minimizing Empirical Risk

- We can solve for *a* in term of *b* along with the means of $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

- By definition of standard deviation and covariance, the solution for *b* is

$$\hat{b} = r\frac{SD_y}{SD_x}$$

- We obtain the expression

$$\hat{y} = \bar{y} + rSD_y\frac{(x - \bar{x})}{SD_x}$$

- Note that linear regression tends to reduce the spread of value among the feature variables.
  - For example, if x is three standard deviations above the mean, then $\hat{y}$ is 3r standard deviations above the mean

Linear Regression

Covariance resembles variance except that it applies to two different random variables

$$\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}} \frac{1}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

For datasets $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$ with mean 0, we can summarize the relationships between the expressions

Correlation measures the cosine of the angle between the dataset thought of as vectors

$$= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \frac{\sqrt{\sum_{i=1}^{n} y_i^2}}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

Standard deviation measures the spread of values around the mean

Decomposing Variance

- Remember that we have discussed decomposing risk and empirical risk for the square loss into two component called
  - Bias measuring the accuracy of an estimator / estimate
  - Variance measuring the consistency of an estimator / estimate
- We can use the same calculation to decompose the variance of the response variable

Expanding the square gives another term. However the term equals 0 because the summation of residuals is 0

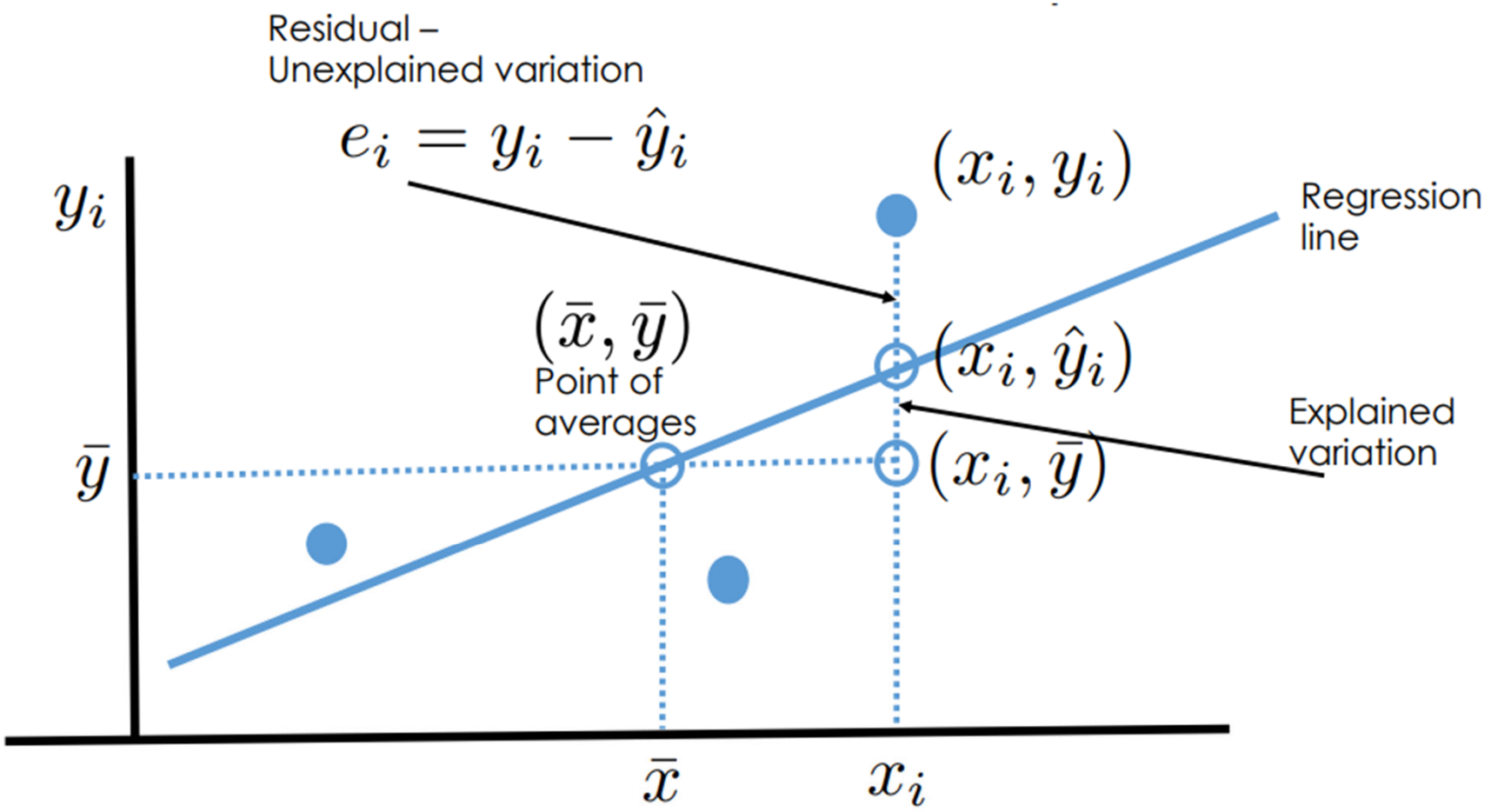
$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \sum_{i}(y_i - \hat{y}_i)^2 + \sum_{i}(\hat{y}_i - \bar{y})^2$$

Unexplained Variance

Explained Variance

Decomposing Variance



Example

Cheap, light, and versatile, plastics are the dominant materials of our modern economy. Their production is expected to double over the next two decades. Yet only 14% of all plastic packaging is collected for recycling after use and vast quantities escape into the environment. This not only results in a loss of USD 80 to 120 billion per year, but if the current trend continues, there could be more plastic than fish (by weight) in the ocean by 2050.

Example

SUNDAY, OCTOBER 20, 1996 San Francisco Examiner

Stripping the Seas

Fishermen are taking from the oceans faster than species can be replenished

By Jane Kay

Commercial fishing vessels are hauling so much seafood from the world's oceans that more than 100 species are in danger, scientists say.

Last week, an international science group issued a "red list" of imperiled marine fish, including some species of tuna, swordfish, shark and Pacific red snapper.

"The ocean cannot sustain the massive removal of wildlife needed to keep nations supplied with present levels of food taken from the sea," said Sylvia Earle, an internationally known marine biologist from Oakland.

Regulators and fishing groups, among them the National Marine Fisheries Service and the Pacific Fishery Management Council, consider the list an indicator of failing fish populations.

The World Conservation Union's 1996 list of 5,205 threatened animals includes 120 marine fish - a record number. It was based on the recommendations of 30 experts brought together by the Zoological Society of London.

Until now, the group has put only a handful of marine fish on the list, which began in the 1960s and is updated every three years. ...

The problems associated with overfishing began more than a half century ago when the annual catch of sea creatures was about 20 million tons.

In the 1940s, demand for fresh fish grew.

With advanced fish-locating technology, factory-size vessels, longer fishing lines and nets, the world catch exceeded 60 million tons by the 1960s. It peaked in 1989 at 86 million tons, then began to dwindle.

In 1993, the National Marine Fisheries Service announced that of 157 commercially valuable fish species in the United States, 36 percent were overfished and 44 percent were fished at the maximum level. ...

Premolt	113.6	118.1	142.3	125.1	98.2	119.5	116.2
Postmolt	127.7	133.2	154.8	142.5	120.0	134.1	133.8
Increment	14.1	15.1	12.5	17.4	21.8	14.6	17.6
Year	NA	NA	81	82	82	92	92
Source	0	0	1	1	1	1	1

Variable	Description
Premolt	Size of the carapace before molting.
Postmolt	Size of the carapace after molting.
Increment	postmolt–premolt.
Year	Collection year (not provided for recaptured crabs).
Source	1=molted in laboratory; 0=capture–recapture.

Summary

- Correlation
  - Capturing liner relationship
- Linear Regression
  - Prediction unknown data from known data
- Example

Goals

- Use correlation to describe linear relationships between random variables
- Predict a response random variable from a feature random variable with linear regression

Questions

- Questions on Piazza?
  - Please provide your feedback along with questions
- Question for You!

Can you find two dependent random variables with correlation equal to 0?