CS/ENGR M148 L3: Simple Linear Regression

Sandra Batista

Administrative News

If you are still forming teams or modified your team, we will open a new form for you to state interests and reopen the team contract assignment**. Announcement will be made on BruinLearn.**

This week in discussion section:

Lab on simple regression

Project Data Check-in: Your team will need to demonstrate some data cleaning and EDA. How can you use EDA to help you plan for prediction and choose variables for simple linear regression?

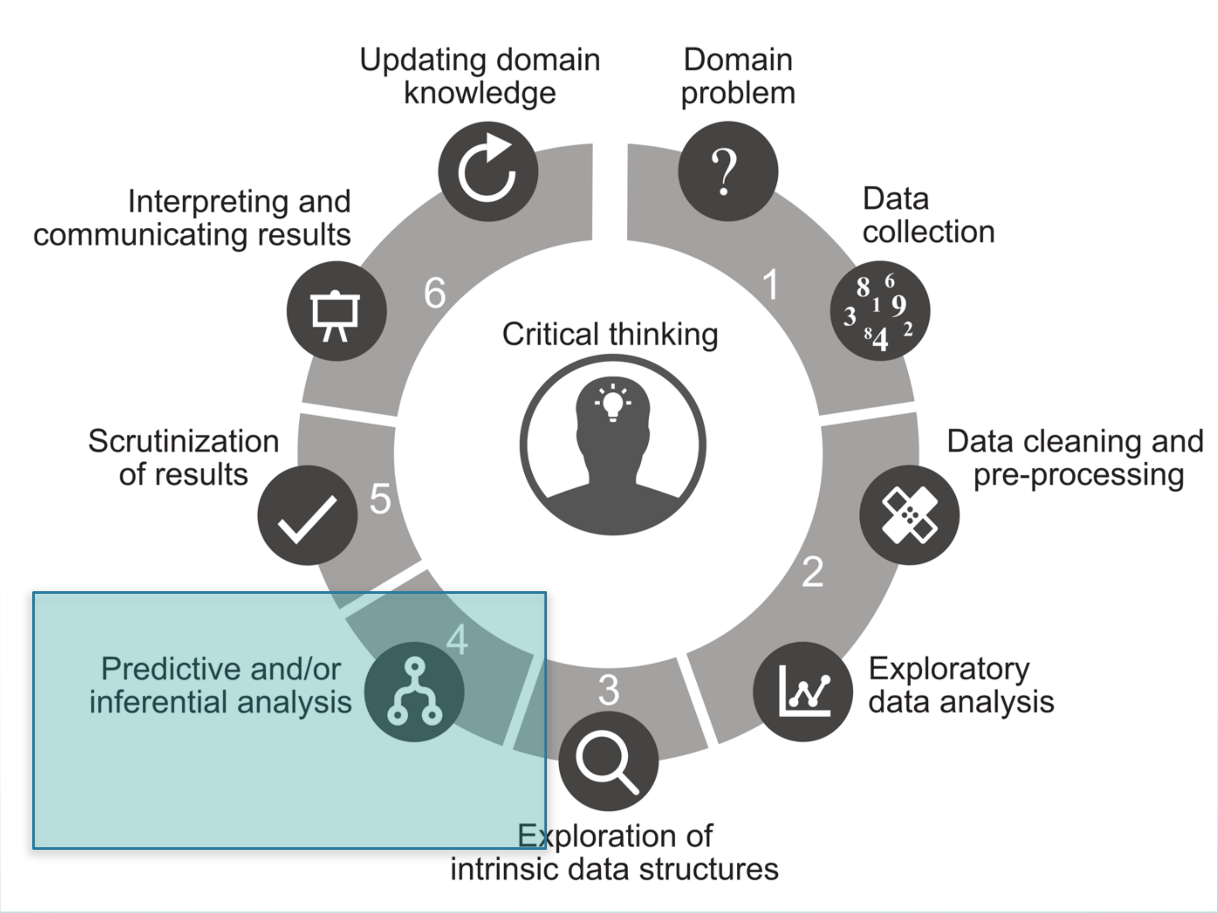
Projects

1. Projects will be graded on how well they demonstrate mastery of the methods taught in class and discussions.
2. You may choose your own data set or a data set supported by the course staff.
3. Team contract 5% - This week during discussion. A sample contract will be made available. Team contracts due by 11:59 pm PT on 10/4/24
4. Project discussion check-ins: 30%, 6x5%
5. Final project code: 25%
6. Final project report: 40%

Join our slido for the week…

https://app.sli.do/event/kJ89kkneBvwrBoxTVCpmcK

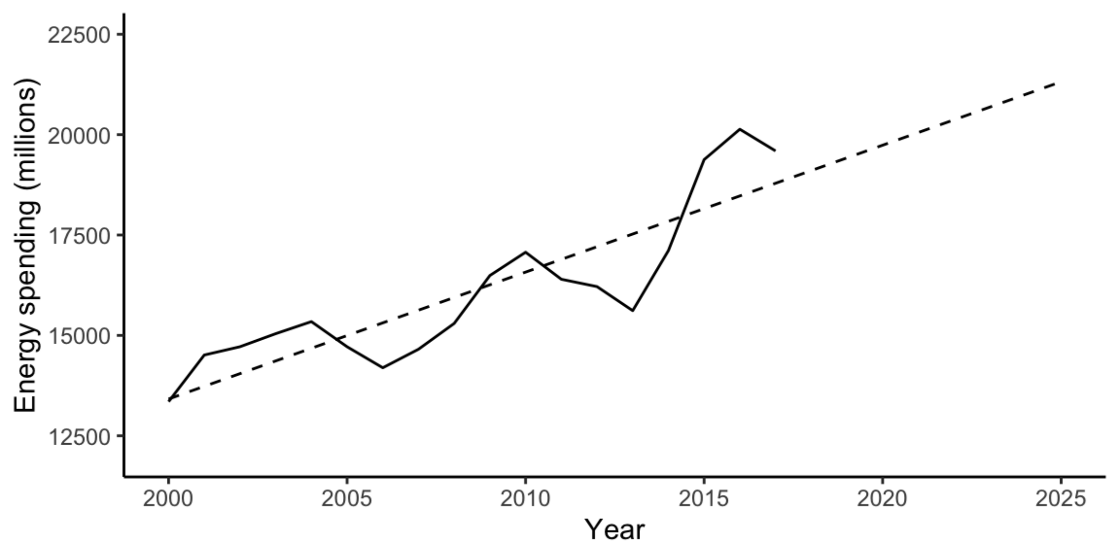
Data Science Life Cycle (DSLC)



[Yu, Barter 2024]

DSLC Step 4: Predictive Analysis

In **prediction problems** our goal is to use past or current observable data to predict something about future unseen data.

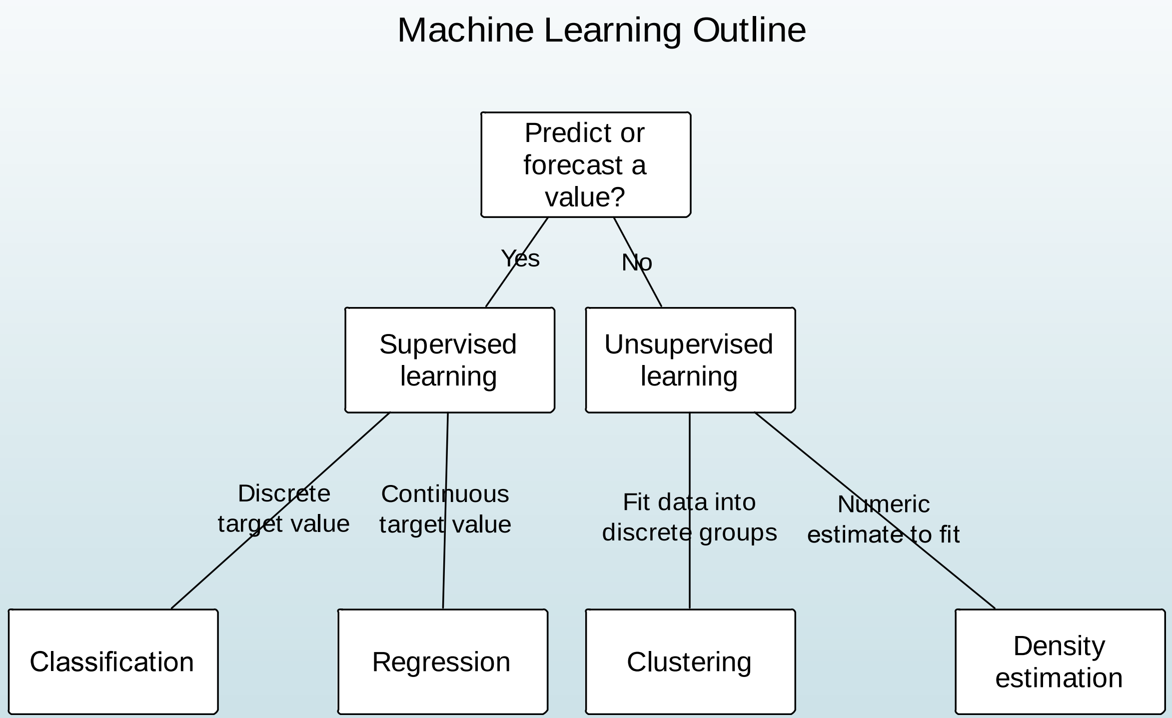


Machine learning methods for prediction include **classification and regression.**

The techniques used are **supervised learning algorithms.**

[Yu, Barter 2024]

Outline of core ML problems



Regression: A process for modeling the relationship between variables of interest

[Shah, 2020]

Data mining

- Understanding the nature of the data to gain insight into the problem that generated the data set in the first place.
- Can be performed by a human expert on a specific data set, often with a clear end goal in mind.

[Shah, 2020]

Today's Learning Objectives

Students will be able to:

- Identify **predictive problems**
- Plan for **prediction** using **sampling**
- Fit **linear models** using **L1 and L2 loss functions**
- Begin exploring the relationship between **least squares and correlation**

6538641

The Modeling Process

1. Choose a model	*How should we represent the world?*
2. Choose a loss function	How do we quantify prediction error?
3. Fit the model	How do we choose the best parameters of our model given our data?
4. Evaluate model performance	How do we evaluate whether this process gave rise to a good model?

Models

A **model** is a some mathematical rule or function to describe the relationships between variables.

Dataset

x	y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

Observation

$$(x_i, y_i)$$

- Independent variable
- **Input**
- **Feature**
- **Attribute**

- Dependent variable
- **Output**
- **Outcome**
- **Response**

Prediction

If we use x to predict y, the predictions are denoted as $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$

Prediction problems

- The goal of a **prediction problem** is to predict the value of a response variable whose value is *unobserved* in future data.
- The variable being predicted is called the **response variable.**
- Variables used in the model to create the prediction are called **predictor variables (predictors, predictive features, covariates, or attributes)**
- We use *observed* response values and the predictive features to create a relationship to generate predictions of the unobserved responses in future data.

[Yu, Barter 2024]

Prediction algorithms

- A **predictive algorithm** aims to predict the value of a response variable based on the values of predictor variables (also known as covariates or predictive features).
- **Predictive algorithms** typically work by finding some particular combination of the predictor variables such that their combined value is as close as possible to the actual response value.
- Assuming relationship in observed data holds in future, the algorithm can be used to predict future values
- Today we'll focus on Least Squares Algorithm

[Yu, Barter 2024]

Define Response Variable

- A **response variable** is value you want to predict such as patient's blood oxygen level from pulse oximeter
- **Labeled data** is needed to train and evaluate a predictive algorithm. **Labeled data** is data where response variable is known.
- **Binary responses** always have one of two possible values, e.g. whether an email is spam (spam/not spam). Used for **classification** problems
- **Continuous responses** can be an arbitrary *numeric* value, e.g., a company's annual revenue (in dollars). Continuous response prediction problems are often called **regression problems**

[Yu, Barter 2024]

Predicting house prices

Examining data from houses sold in Ames, IA from 2006 to 2010 that has been provided by De Cock (2011) from the Ames City Assessor's Office.

Response variable: home price

What should predictors be?

[Yu, Barter 2024]

Your turn:
Predicting house prices

Please get the Jupyter notebook

Go to:
https://colab.research.google.com/drive/16HKs6Nz4UBvhil5912oWdqXuSowTCp8f?usp=sharing
We'll learn about the data and load it...

Save a copy to your Google Drive and keep notes there...

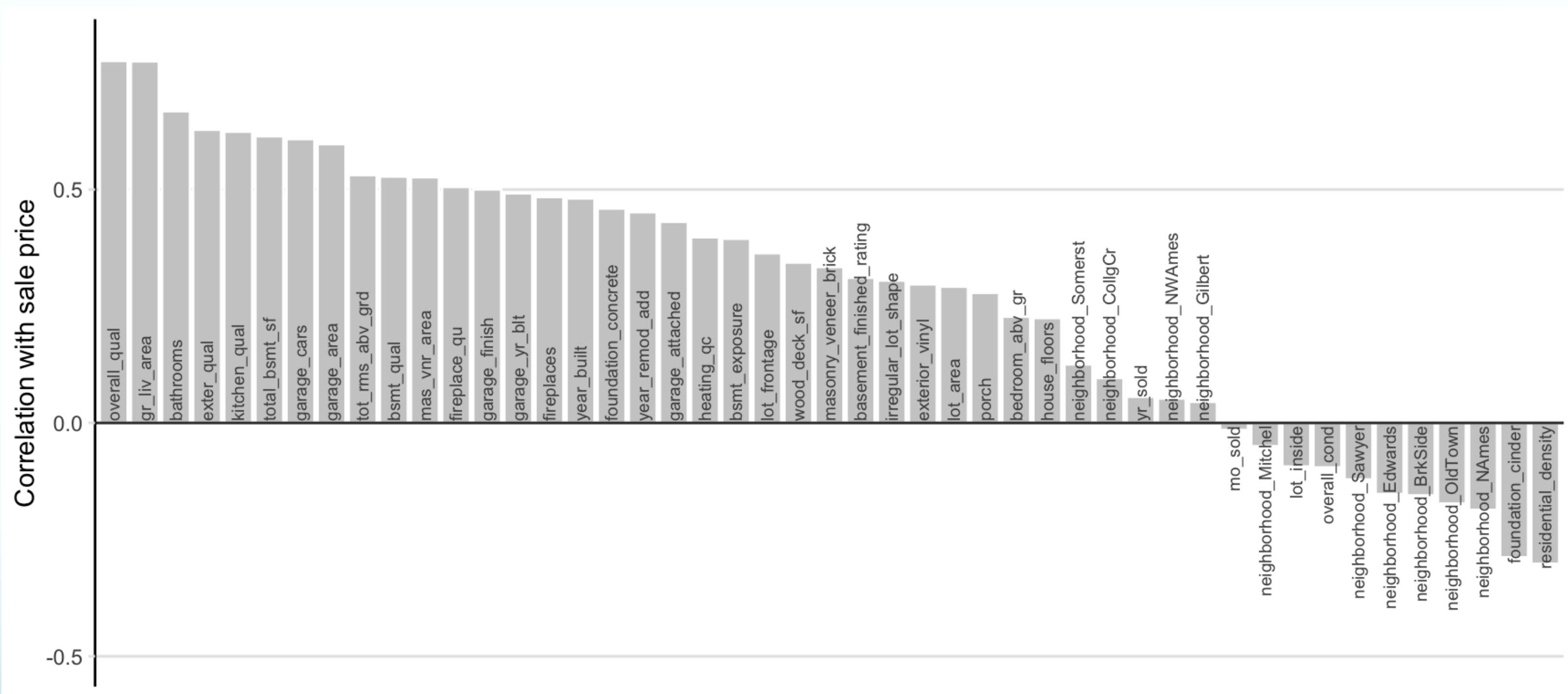
[Yu, Barter 2024]

Define Predictors

- How do you know what variables to use as predictors?
- Example: predicting the sale price of a house (the response variable predictive features: house size, age, condition
- **Use domain knowledge**
- **Use EDA** to find a small set of high quality predictors

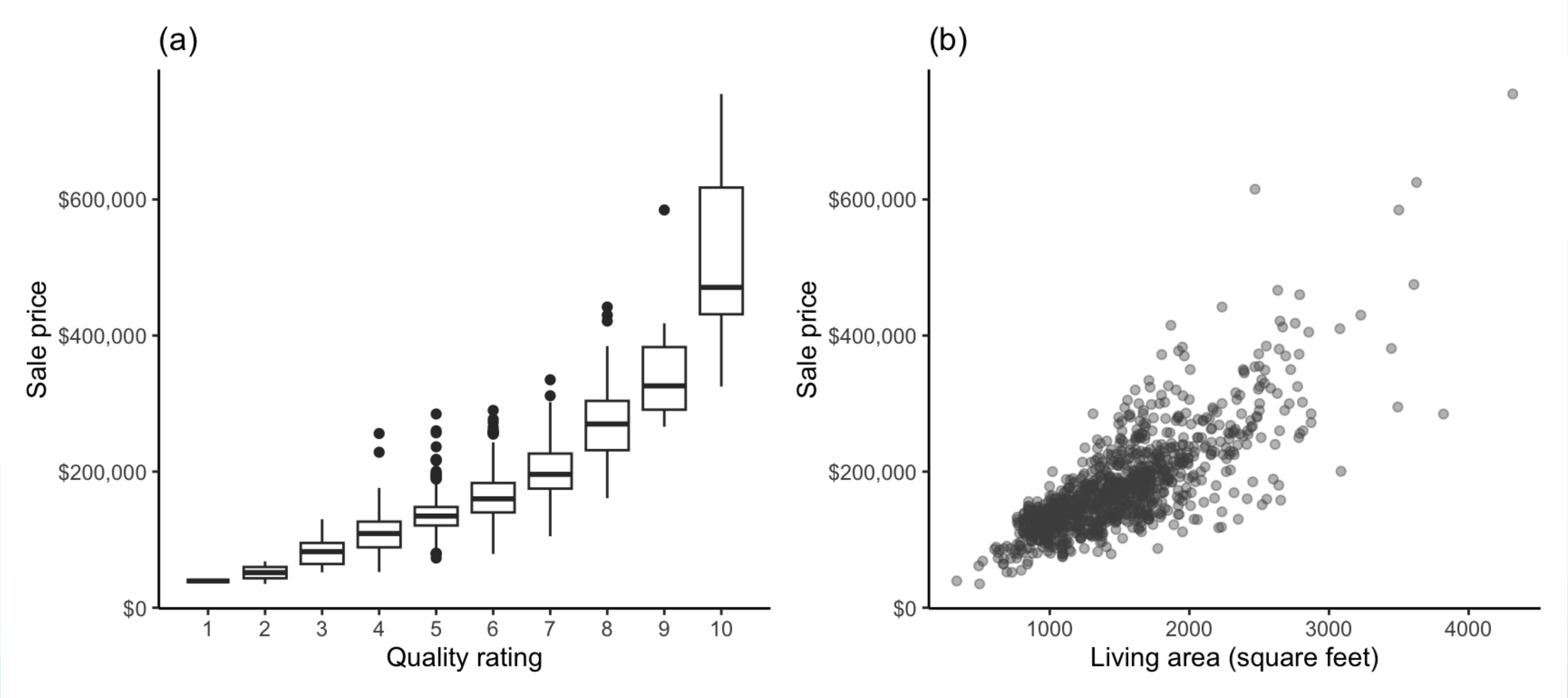
[Yu, Barter 2024]

EDA for house prices



[Yu, Barter 2024]

EDA for house prices



[Yu, Barter 2024]

Today’s Learning Objectives

Students will be able to:

- ✓ Identify **predictive problems**
- ✗ Plan for **prediction** using **sampling**
- ✗ Fit **linear models** using **L1 and L2 loss functions**
- ✗ Begin exploring the relationship between **least squares and correlation**

Plan for Predictability

- How do you know what variables to use as predictors?
- Example: predicting the sale price of a house (the response variable predictive features: house size, age, condition
- **Use domain knowledge**
- **Use EDA** to find a small set of high quality predictors

[Yu, Barter 2024]

Predictability

"Data-driven results are **predictable** if they can be shown to reemerge in (i.e., can be generalized to) new, relevant scenarios"

- This can apply to separate or future data sets
- We need **labeled data to evaluate predictions.**

- A single data set can be partitioned into a **training set (60%), validation set (20%), test set(20%)**

Sets can be partitioned on **time-based splits, group-based splits,** or **randomly.**

Inference (Prediction): drawing conclusions about a population based on a sample.

Sampling

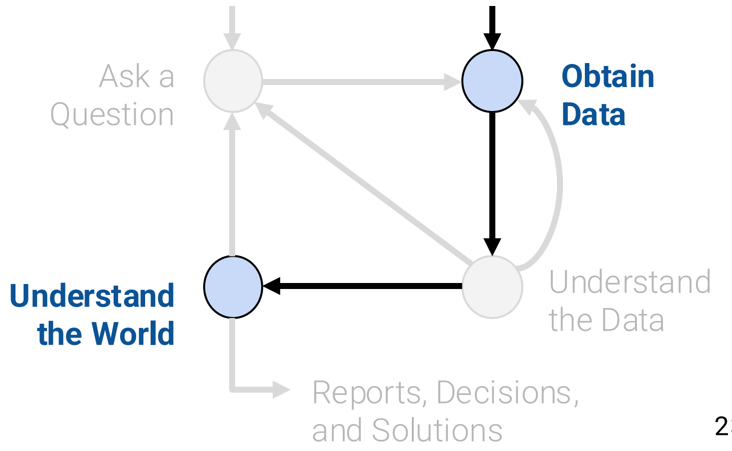
A **sample** is (usually) a subset of the population.

- Samples are often used to make **inferences about the population**.

- How you draw the sample will affect your accuracy.

Sources of error:

- **chance error**: random samples can vary from what is expected, in any direction.
- **bias**: a systematic error in one direction.
  - Could come from our sampling scheme and survey methods.



6538641

Probability Sample (aka Random Sample)

Why sample at random?

1. To get more representative samples → **reduce bias**
    - However, the **choice of randomization** can still introduce bias.
2. More importantly, with random samples we can **estimate** the **bias** and **chance error** → **quantify uncertainty**

For a **probability sample**,

- We have to be able to provide the **chance** that any specified **set** of individuals will be in the sample.
- All individuals in the population **need not** have the same chance of being selected.
- Because we know all the probabilities, we will be able to **estimate the errors**.

Inference (Prediction): drawing conclusions about a population based on a sample.

Common Biases

Selection Bias

- Systematically excluding (or favoring) particular groups.
- **Example**: Medical study on pulse oximeters only used cohort of patients from predominately one population
- **How to avoid**: Examine the sampling frame and the method of sampling.

Response Bias

- People don't always respond truthfully, or questions lead to certain responses.
- **Example**: Patients may not recall some of their measurements correctly.
- **How to avoid**: Examine the nature of questions and the method of surveying.
    - **Randomized Response** – flip a coin answer yes if heads or truthfully if tails.

Non-response Bias

- People don't always respond → People who don't respond aren't like the people who do!
- **Example**: Some patients may not reply to survey at all if too time consuming or invasive.
- **How to avoid**: Keep your surveys short, and be persistent.

Inference (Prediction): drawing conclusions about a population based on a sample.

Convenience Samples

An example of a non-random sample is ***convenience sample.*** *It's whatever we can get ahold of.*

Example: Scientists in New South Wales (AUS) collect specimens from eucalyptus trees to keep in museums, recording **where they came from** in latitude / longitude.

Can we use this data to map the ***geographic distribution*** *of eucalyptus trees?*

Warning:

- *Haphazard ≠* ***random****.*
- *Many potential sources of bias!*

6538641

Common random sampling schemes

A **simple random sample (SRS)** is a sample drawn **uniformly** at random **without** replacement.

- **Every individual (and subset of individuals) has the same chance of being selected.**
- Every pair has the same chance as every other pair.
- Every triple has the same chance as every other triple.
- And so on.

A **uniform random sample with replacement** is a sample drawn **uniformly** at random **with** replacement.

- Similar to SRS but some individuals in the population might get picked more than once.
- **Easier to compute probabilities then SRS**
  - Approximation of large SRS
- Not used in practice because surveying someone twice is wasteful

A raffle could use either sampling scheme, depending on if winners are eligible for multiple prizes.

Your turn: Predicting house prices

Please get the Jupyter notebook

Go to:
https://colab.research.google.com/drive/16HKs6Nz4UBvhiI5912oWdqXuSowTCp8f?usp=sharing

We'll learn about the data and load it...

Save a copy to your Google Drive and keep notes there...

[Yu, Barter 2024]

Random Sampling Code

```
## this code would define the training, validation, and test set equivalent

 ames_train = ames.query("`Mo Sold` <= @split_date_month & `Yr Sold` <=
@split_date_year")

## filter to houses not in training set

 ames_val = ames.query("~PID.isin(@ames_train.PID)")

## randomly select half of the houses for the validation set

ames_val = ames_val.sample(round(len(ames_val.index)*0.5), random_state=3789)

## filter to houses not in training and validation sets for the test set

 ames_test = ames.query("~PID.isin(@ames_train.PID) & ~PID.isin(@ames_val.PID)")
```

Today's Learning Objectives

Students will be able to:

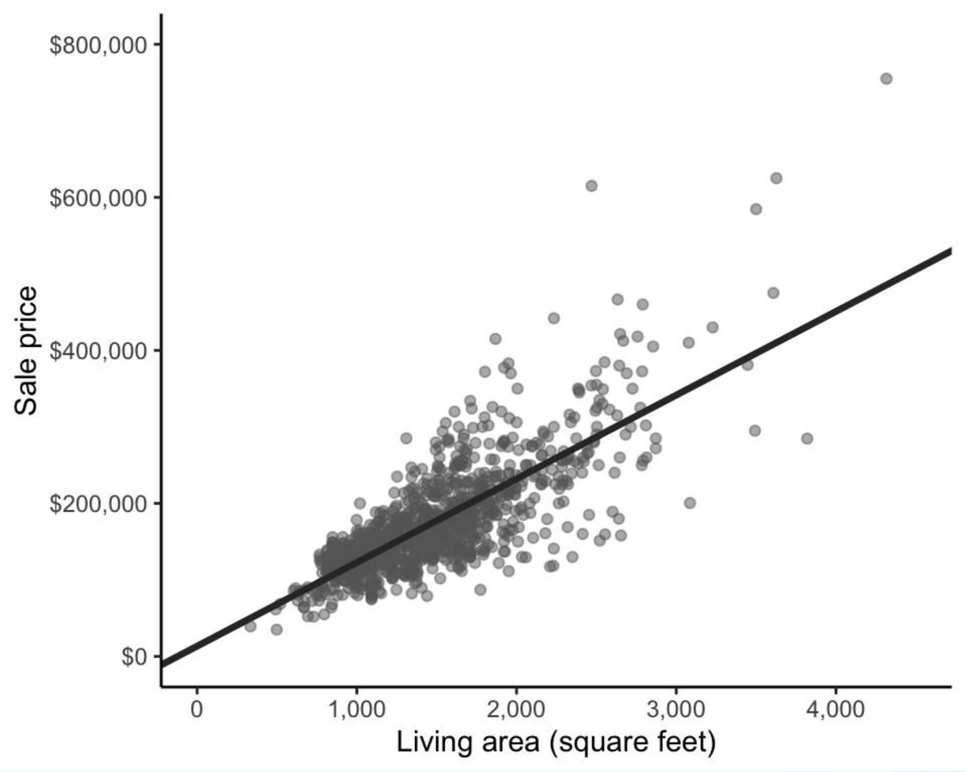
- ✓ Identify **predictive problems**
- ✓ Plan for **prediction** using **sampling**
- Fit **linear models** using **L1 and L2 loss functions**
- Begin exploring the relationship between **least squares and correlation**

Visualize Predictive Relationship

Fitted line for linear relationship but is it "best"?

Caution: Are there any **confounders?**

A **confounder** is a common cause of increases in both size and price.



[Yu, Barter 2024]

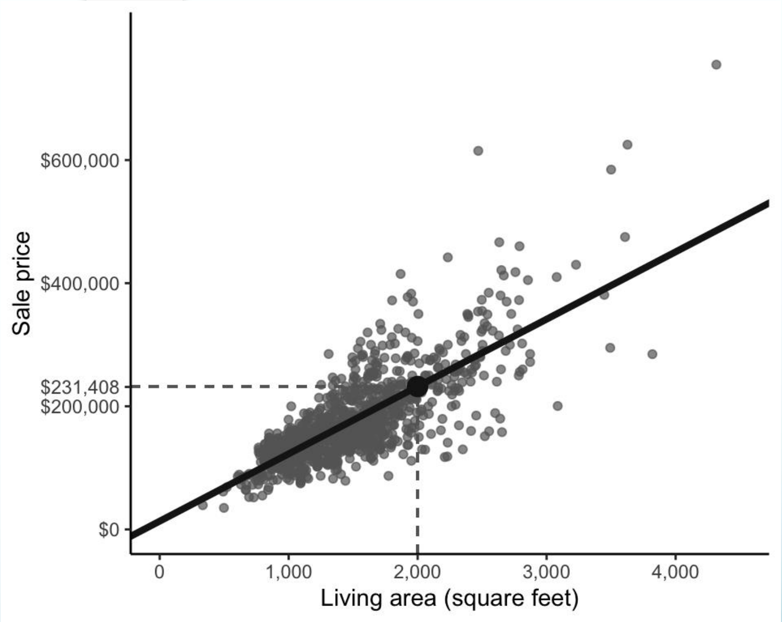
Fitted Line for prediction

$$predicted\,price = b\_0 + b\_1 \times area.$$

b0 = 13,408 is **the intercept** of the line

b1 = 109 is the **coefficient** of

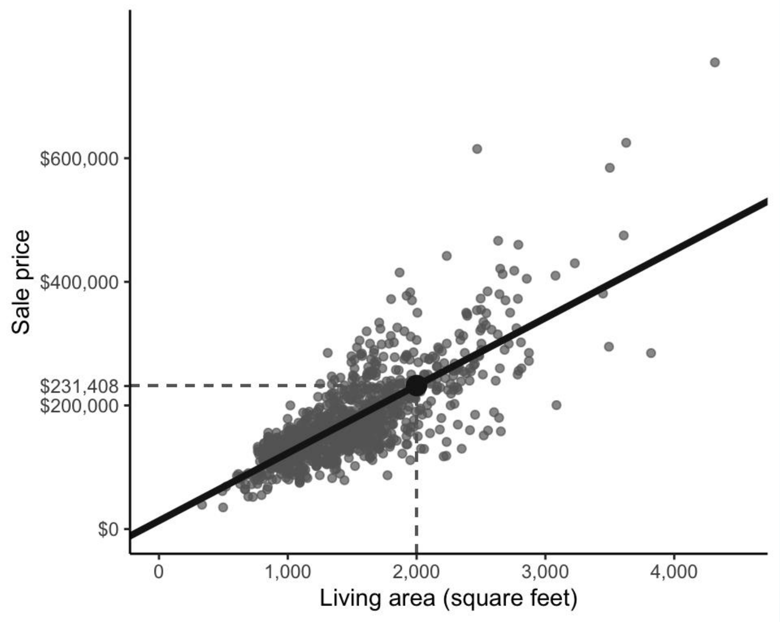
the predictor variable



[Yu, Barter 2024]

What fit is best?

We want to *minimize* an **objective** function or the **loss** function



The "loss" measures

what you lose when you

make the prediction

(i.e.,

how different your predictions are from the observed values).

[Yu, Barter 2024]

Least Absolute Deviation

L1 loss function or absolute value loss function:

$$\frac{1}{n}\sum_{i=1}^{n}\left|\text{observed response}_i - \text{predicted response}_i\right|,$$

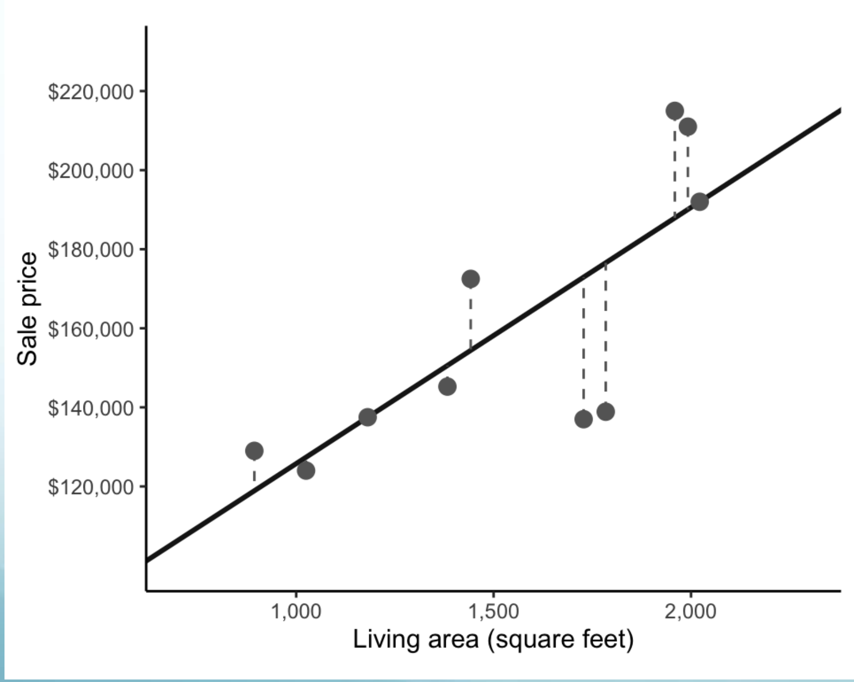
$$\frac{1}{n}\sum_{i=1}^{n}\left|\text{observed price}_i - (b_0 + b_1 \times \text{area}_i)\right|.$$

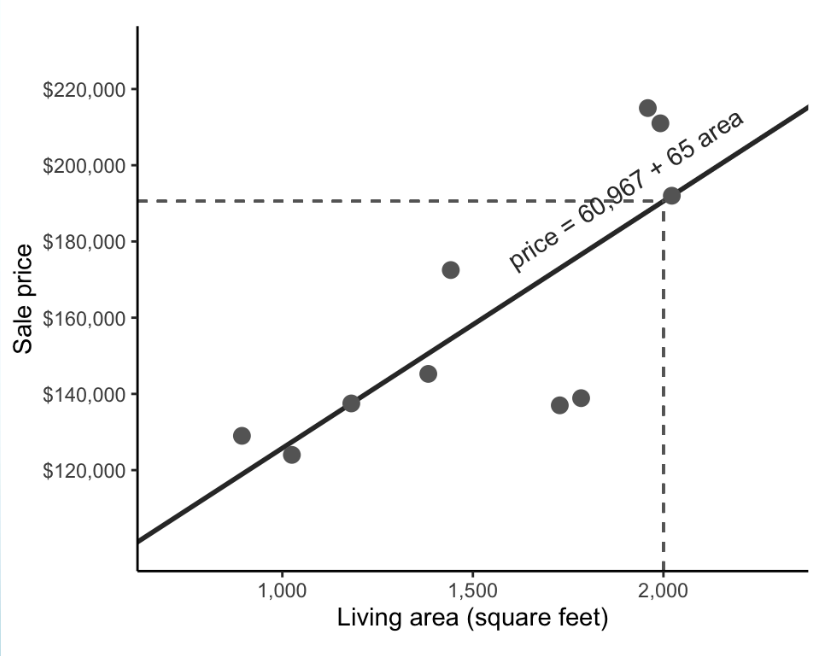
$$\frac{1}{n}\sum_{i=1}^{n}\left|y_i - (b_0 + b_1 x_i)\right|.$$

[Yu, Barter 2024]

Least Absolute Deviation

The Least Absolute Deviation (LAD) algorithm generates a fitted line to minimize the absolute value (or L1) loss function





[Yu, Barter 2024]

Your turn: Predicting house prices

Please get the Jupyter notebook

Go to:
https://colab.research.google.com/drive/16HKs6Nz4UBvhiI5912oWdqXuSowTCp8f?usp=sharing

We'll learn about the data and load it...

Save a copy to your Google Drive and keep notes there...

[Yu, Barter 2024]

Least Squares Loss Function

L2 loss function or squared loss:

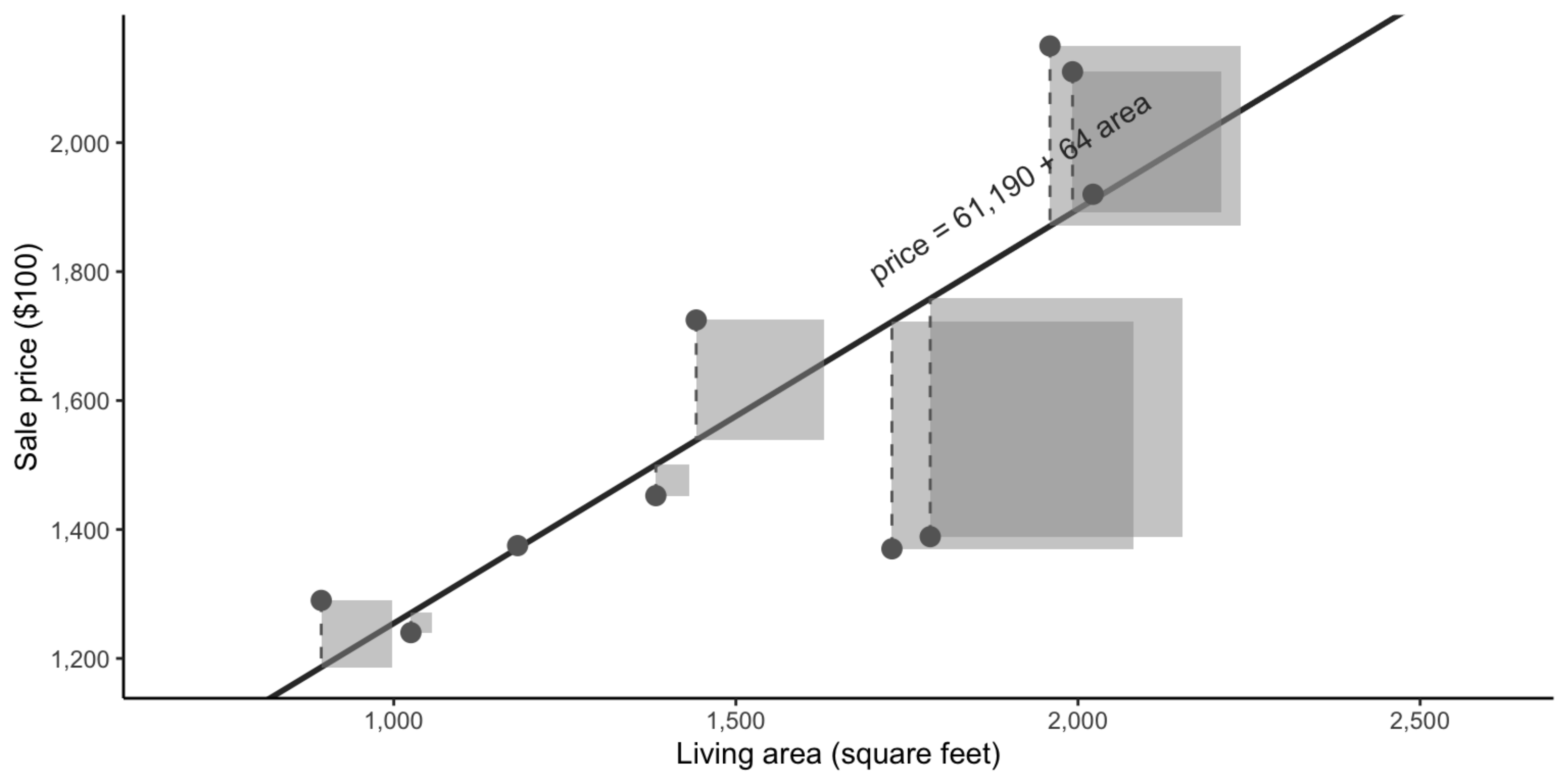
$$\frac{1}{n}\sum_{i=1}^{n}(\text{observed response}_i - \text{predicted response}_i)^2.$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2.$$

[Yu, Barter 2024]

Least Squares

The **Least Squares (LS) algorithm** generates a fitted line by minimizing the squared (or L2) loss function



[Yu, Barter 2024]

Closed Form for Estimates

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$b_0 = \overline{\text{observed price}} - b_1 \times \overline{\text{area}} \text{ and}$$

$$b_1 = \frac{\sum_{i=1}^{10}(\text{area}_i - \overline{\text{area}})(\text{observed price}_i - \overline{\text{observed price}})}{\sum_{i=1}^{10}(\text{area}_i - \overline{\text{area}})^2},$$

$$\text{predicted price} = 61,190 + 64 \times \text{area}.$$

[Yu, Barter 2024]

Your turn: Predicting house prices

Please get the Jupyter notebook

Go to:
https://colab.research.google.com/drive/16HKs6Nz4UBvhiI5912oWdqXuSowTCp8f?usp=sharing

We'll learn about the data and load it…

Save a copy to your Google Drive and keep notes there…

[Yu, Barter 2024]

Today's Learning Objectives

Students will be able to:

- ✓ Identify **predictive problems**
- ✓ Plan for **prediction** using **sampling**
- ✓ Fit **linear models** using **L1 and L2 loss functions**
- ✗ Begin exploring the relationship between **least squares and correlation**

Statistical representation of data

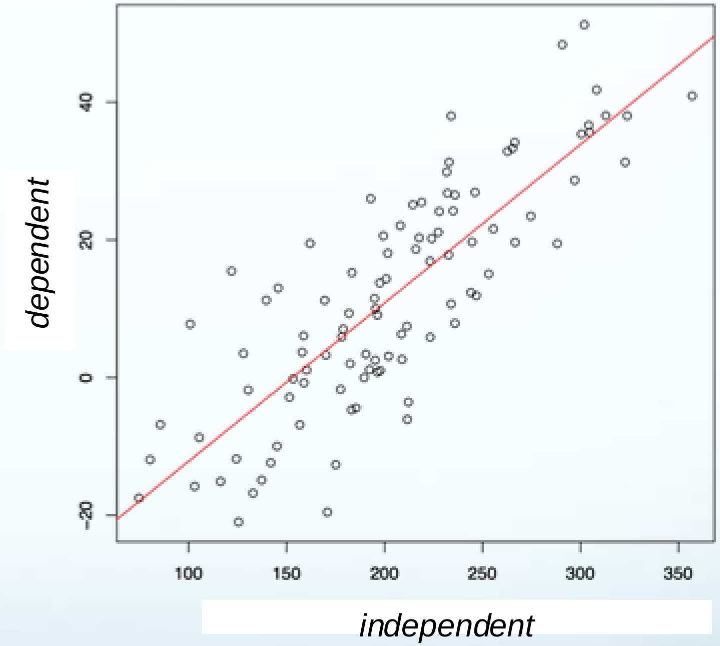
- The convention for representing data tables in ***statistics*** is to use the *rows* for observations, and the *columns* for features. Moreover, *n* is used to represent the number of observations and *p* the number of features, so that a data table has size *n x p*.
  - One reason for this convention is the form of regression models, which describe observations as linear combinations of explanatory variables with some added noise using the form:
  - With this matrix notation, *X*, which is also known as the design matrix, has dimensions *n x p*.

$$y = X\beta + \epsilon.$$

Linear regression

- "Regression analysis" refers to the problem of estimating relationships between a dependent variable, and one or more independent variables.

- The simplest example of this is linear regression, where the relationship to is assumed to be linear, and regression analysis then refers to finding a linear combination of the independent variables that provides the *best fit* to the dependent variable.

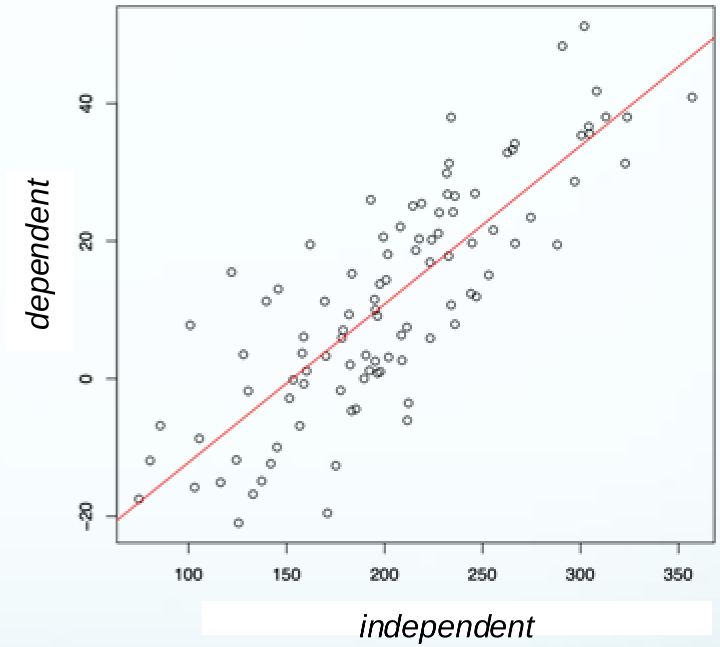


Least squares

- Fit a line of the form *y=mx+b*.
- **Goal:** find a line with the property that the average (vertical) <u>loss</u> between the points and the line is minimized.
- use squared distance for the <u>loss function</u> because its optimization is easier than the alternatives.

$$r_i = y_i - f(x_i, \boldsymbol{\beta}).$$

$$S = \sum_{i=1}^{n} r_i^2.$$



Solving the least squares problem

- Method 1: (Multivariable) calculus.

$$f(m,b) = \sum_{i=1}^{n} (y_i - (mx_i + b))^2.$$

- The task is to minimize *f(m,b)* over the parameters *m* and *b*. The square is helpful because the derivatives of *f* with respect to *m* and *b* are linear.

- Compute the two partial derivatives (with respect to *m* and *b*), set them equal to 0, …

Solving the least squares problem

- Method 2: Linear algebra.
- Consider a column vector formed from the dependent variables, i.e. $Y = (y_1,\ldots,y_n)^T$, as a point in an n-dimensional vector space. Observe that the least squares optimization problem is equivalent to finding the nearest point on a subspace spanned by a column matrix X defined from the dependent variables. Formally, we seek to find the value β that minimizes $(||X\beta - Y||_2)^2$; the minimal β is denoted $\hat{\beta}$.

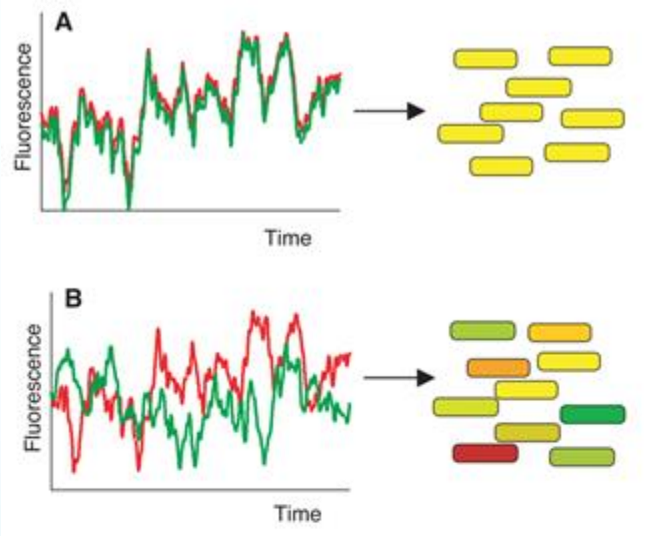
- Using orthogonality, the solution emerges naturally as $(X^TX)^{-1}X^TY$.

Zero-dimensional regression

- One way to think about least squares: generalization of the mean.
- Consider the problem of finding the “closest” number to a set of number $x_1,\ldots x_n$. By “closest” we mean in the sense of squared difference:
- $$\sum_{i=1}^{n}(m - x_i)^2$$
  is minimized. A straightforward (calculus) calculation shows that the minimum is achieved at the mean, i.e.
  $$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$
- Least squares can be viewed as an extension of the mean to higher dimensions.

Stochastic gene expression in a single cell

- In *E. coli*, two reporter genes with same promoters but distinguishable alleles of green fluorescent protein
- In panel A, there appears to be **"correlated" fluctuation** of the two fluorescent proteins.
- In panel B, there appears to be **"uncorrelated" fluctuation** of the two fluorescent proteins.

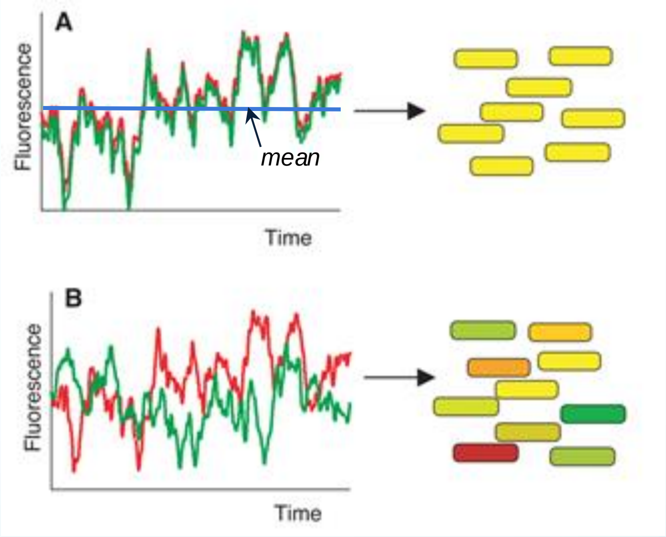


Elowitz et al., 2002

Stochastic gene expression in a single cell

- The fluctuations may be seen as variance, but variance has precise definition in statistics
- The variance of a random variable *X* is

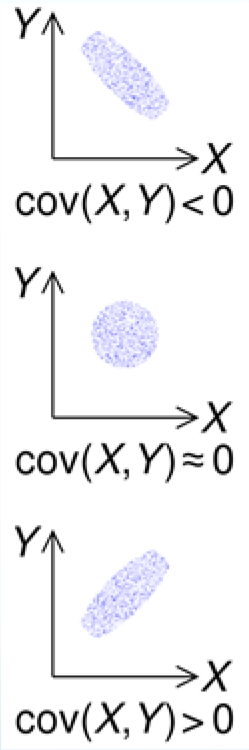
$$Var[X] = E[(X-E[X])^2]$$
$$= E[X^2]-E[X]^2.$$



Elowitz et al., 2002

Covariance of random variables

- The covariance of two random variables X and Y is
  $cov[X,Y] = E[\,(X-E[X])(Y-E[Y])\,]$
  $= E[XY]-E[X]E[Y]$.
- $cov[X,X] = var[X]$.
- If X and Y are independent random variables, then the covariance is zero:
  Proof: Independence means that $E[XY]=E[X]E[Y]$.
  The converse is not true.



Sample covariance

- The sample covariance formula follows from *cov[X,Y] = E[XY]-E[X]E[Y]* :

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

- For unbiased estimator 1/n is replaced by 1/(n-1)

Correlation

- The covariance of two random variables *X* and *Y* is in units that are a product of those of *X* and *Y*. To obtain a dimensionless number, the covariance can be divided by the product of the standard deviation of *X* and the standard deviation of *Y*. This is called the ***correlation coefficient***:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

  Other names include Pearson's product-moment correlation coefficient, Pearson's coefficient, or Pearson's correlation.

Sample correlation coefficient

- The sample correlation coefficient, denoted by *r*, is an estimate of the (population) Pearson correlation. There are several equivalent expressions; the analogue of the sample covariance formula we used is:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

- Note that while Pearson's correlation coefficient lies between -1 and 1 (inclusive), sampling error will reduce the range of *r*.

What is the relationship between linear regression and correlation?

- Let's think about this and pick up here next time…

Today's Learning Objectives

Students will be able to:

- ✓ Identify **predictive problems**
- ✓ Plan for **prediction** using **sampling**
- ✓ Fit **linear models** using **L1 and L2 loss functions**
- ✓ Begin exploring the relationship between **least squares and correlation**

Citations:

Yu, B., & Barter, R. L. (2024). Veridical data science: The practice of responsible data analysis and decision making. The MIT Press.

Shah. C. (2020) A hands-on introduction to data science. Cambridge University Press.

Data 100, Fall 2024, UC Berkeley.