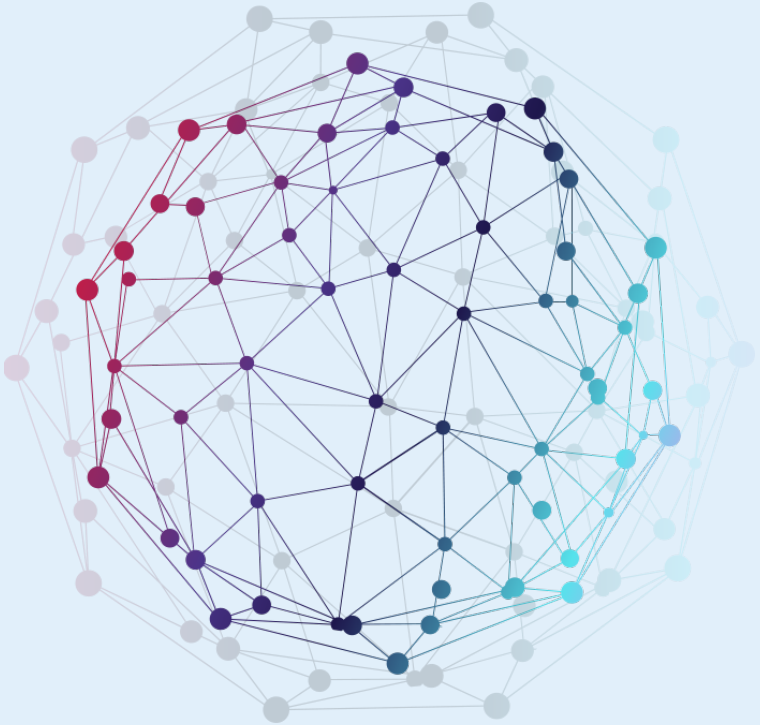


Intro to Statistics and Machine Learning in R

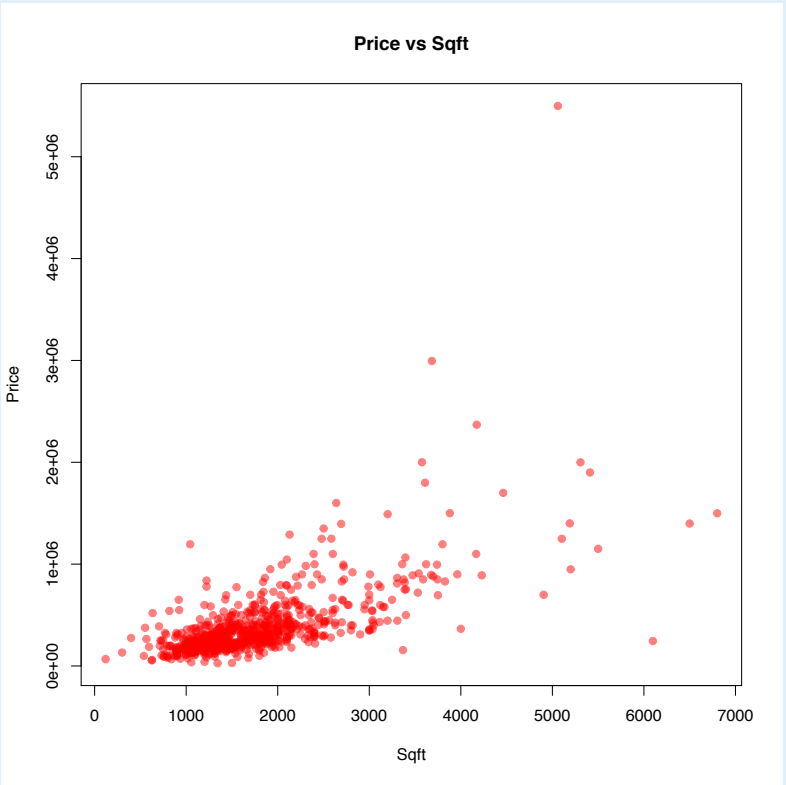
Weiwei Pan

June 13th, 2016

Outline

Motivation

Californian Home Prices (2009)

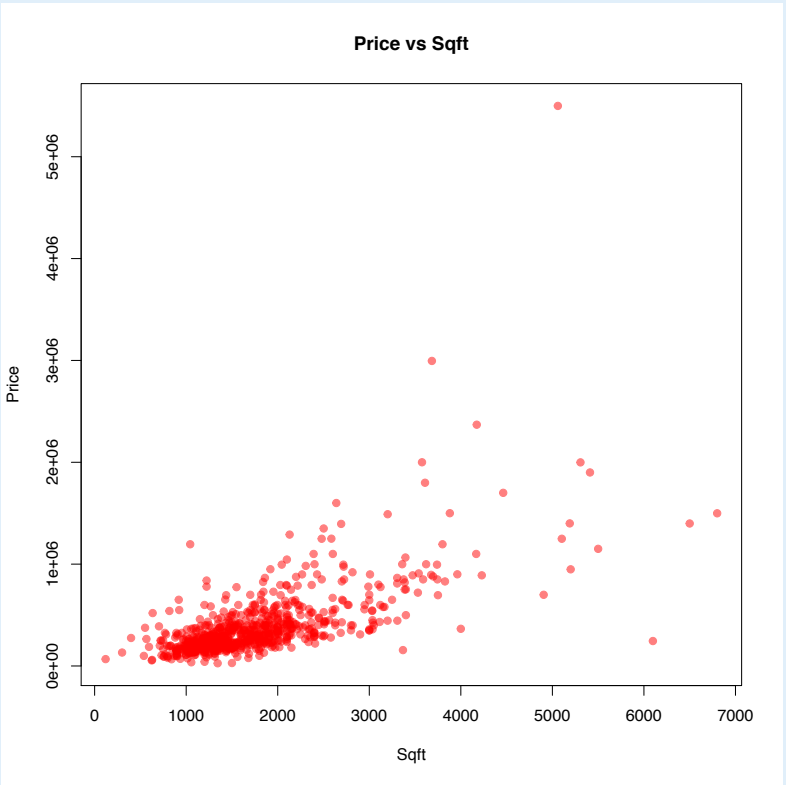


This is a scatter plot of home prices vs square footage of some homes in southern California.

Can you see any patterns or trends?

```
plot(mydata$Sqft, mydata$Price, main="Price vs Sqft", col=rgb(1,0,0,0.5),
xlab="Sqft", ylab="Price", pch=19)
```

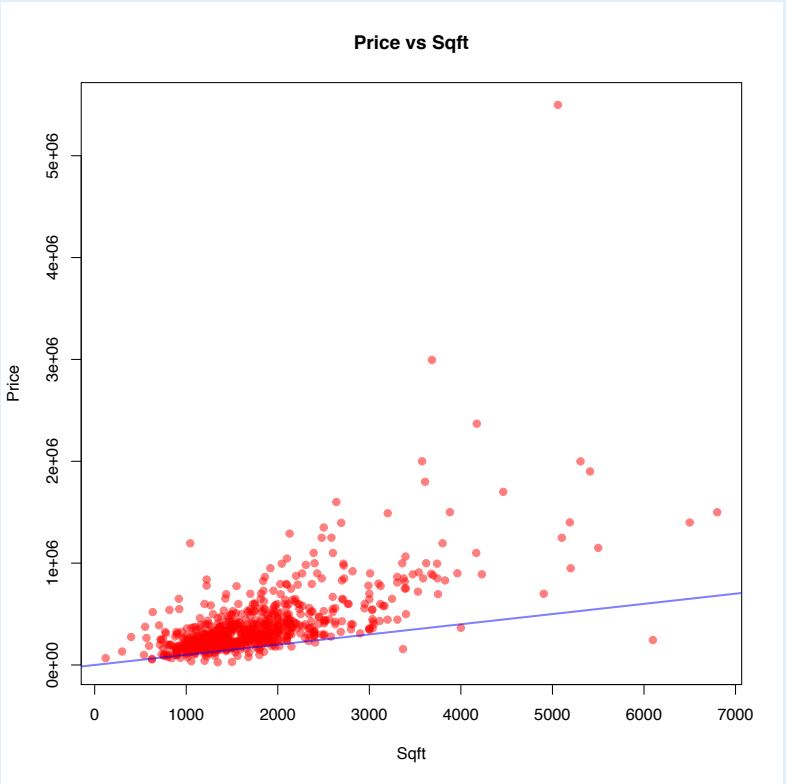
Californian Home Prices (2009)



We see that **as square footage increases, so does price**. But what is a precise, mathematical description of this relationship?

```
plot(mydata$Sqft, mydata$Price, main="Price vs Sqft", col=rgb(1,0,0,0.5),
xlab="Sqft", ylab="Price", pch=19)
```

Californian Home Prices (2009)

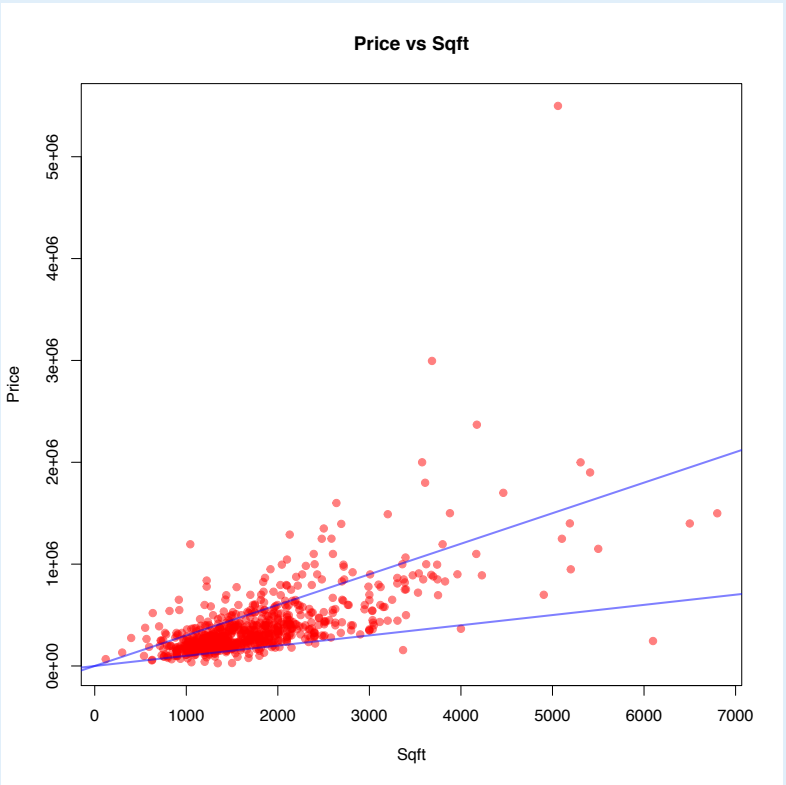


Maybe we want to model the relationship between square footage and price using a simple line.

Does this line capture the trend in the data?

```
plot(mydata$Sqft, mydata$Price, main="Price vs Sqft", col=rgb(1,0,0,0.5),
xlab="Sqft", ylab="Price", pch=19)
abline(a=intercept, b=slope, col=rgb(0,0,1,0.5), lwd=2)
```

Californian Home Prices (2009)

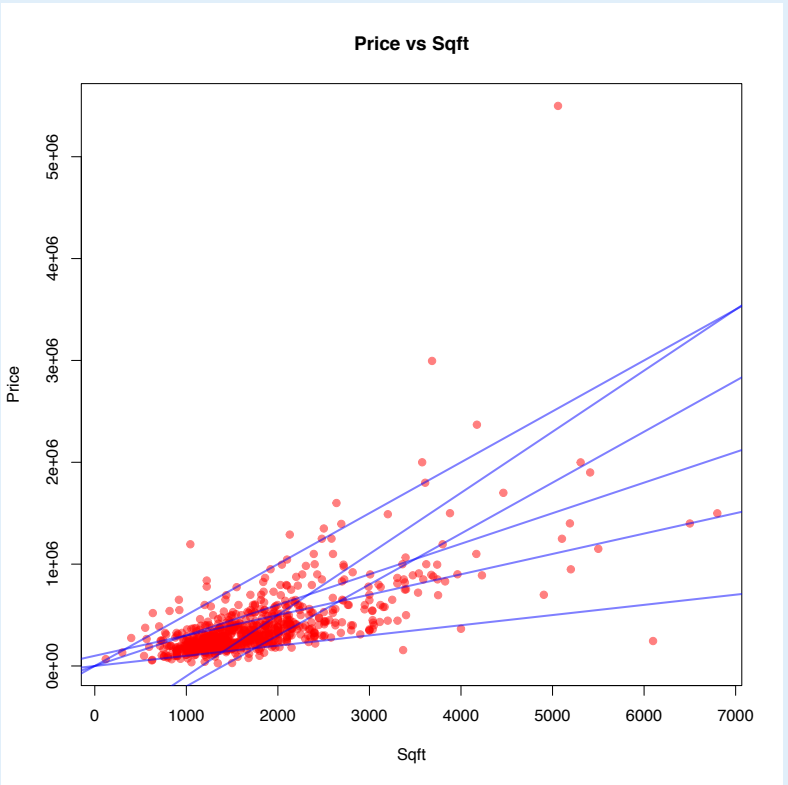


Maybe we want to model the relationship between square footage and price using a simple line.

What about this line?

```
plot(mydata$Sqft, mydata$Price, main="Price vs Sqft", col=rgb(1,0,0,0.5),
xlab="Sqft", ylab="Price", pch=19)
abline(a=intercept, b=slope, col=rgb(0,0,1,0.5), lwd=2)
```

Californian Home Prices (2009)



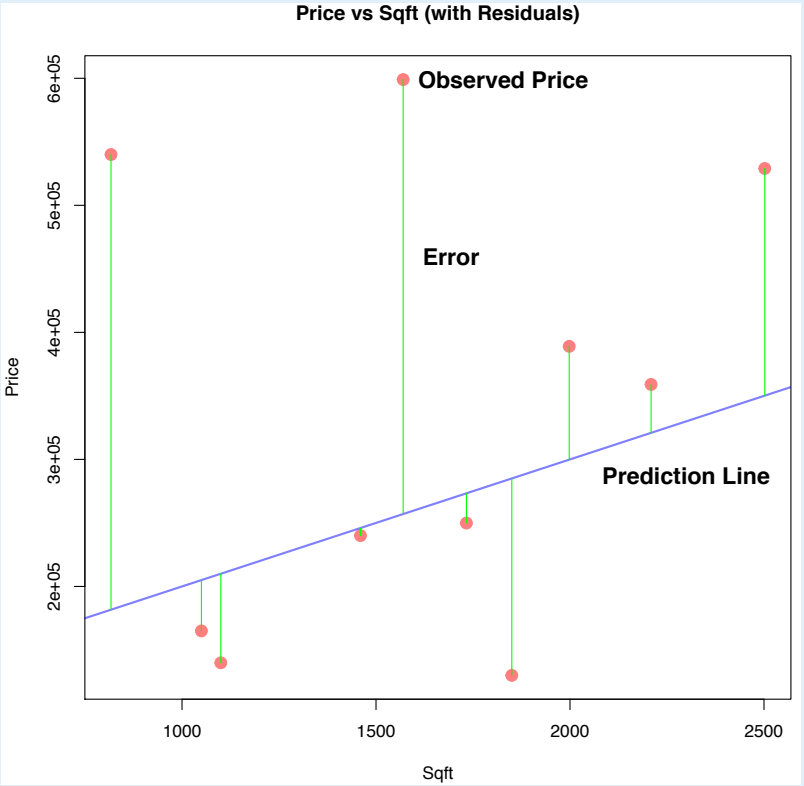
In fact, there are infinite number of lines we can draw through the data.

Which is the best line?

What is a good definition for "*the best line*"?

```
plot(mydata$Sqft, mydata$Price, main="Price vs Sqft", col=rgb(1,0,0,0.5),
xlab="Sqft", ylab="Price", pch=19)
abline(a=intercept, b=slope, col=rgb(0,0,1,0.5), lwd=2)
```

Notion of Error

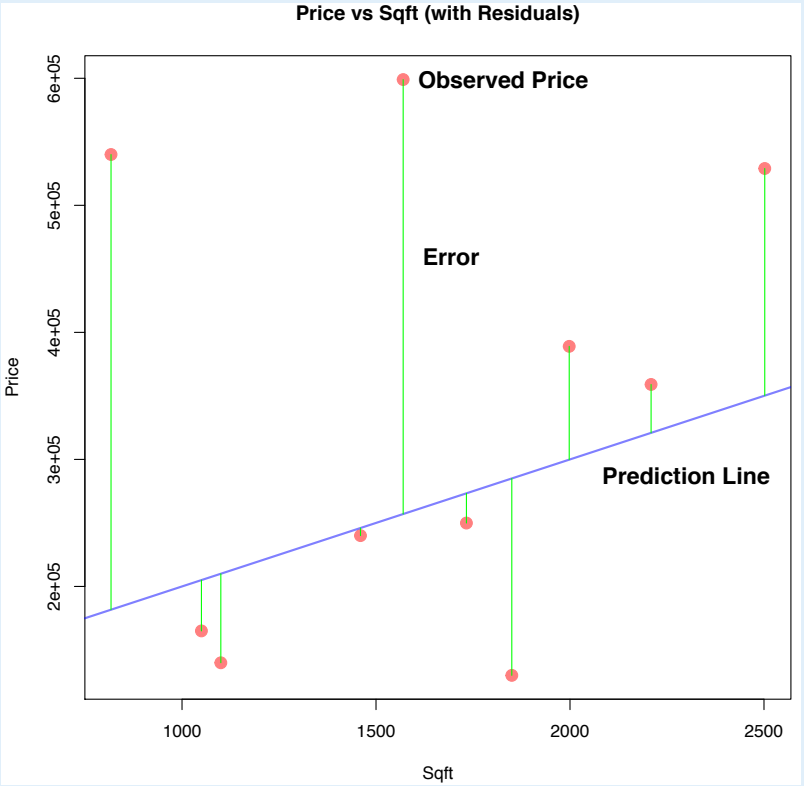


An *absolute residual* is the absolute difference between the actual price of a home and the price predicted by the line for a given square footage.

$$\text{Res}_i = |\text{Observed}_i - \text{Predicted}_i|$$

```
model <- function(x) slope * x + intercept
predict <- sapply(mysample$Sqft, model)
segments(mysample$Sqft, mysample$Price, mysample$Sqft, predict, col="green")
```

Notion of Error



The i-th absolute residual measures the magnitude of the "error" made by the i-th prediction.

```
model <- function(x) slope * x + intercept
predict <- sapply(mysample$Sqft, model)
segments(mysample$Sqft, mysample$Price, mysample$Sqft, predict, col="green")
```

Notions of Fitness

Question: How do we quantify the overall error?

1. (**Max absolute deviation**) Count only the biggest "error"

$$\max_i |\text{Observed}_i - \text{Predicted}_i|$$

2. (**Sum of absolute deviations**) Add up all the "errors"

$$\sum_i |\text{Observed}_i - \text{Predicted}_i|$$

   We can also average them.

3. (**Sum of squared errors**) Add up the squares of the "errors"

$$\sum_i |\text{Observed}_i - \text{Predicted}_i|^2$$

   We can also average them.

Model Fitting

Question: What do we mean by choosing "the best line"?

Answer: A line which minimizes the overall error.

Example: Given a set of points $(x_1, y_1), \ldots, (x_n, y_n)$, the average of absolute deviations of a line $y = mx + b$ is

$$L(m, b) = \frac{1}{n} \sum_{i=1}^{n} |y_i - (mx_i + b)|$$

L is called the lost (or cost) function. Our goal is to **find** $\hat{m}$ **and** $\hat{b}$ **such that the lost,** $L(\hat{m}, \hat{b})$**, is minimal**:

$$(\hat{m}, \hat{b}) = \operatorname{argmin}_{m,b} L(m, b).$$

Finding the optimal values $(\hat{m}, \hat{b})$ is called *fitting the linear model*.

Model Fitting

Question: What do we mean by choosing "the best line"?

Answer: A line which minimizes the overall error.

Example: Given a set of points $(x_1, y_1), \ldots, (x_n, y_n)$, the average of squared deviations of a line $y = mx + b$ is

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

L is called the lost (or cost) function. Our goal is to **find** $\hat{m}$ **and** $\hat{b}$ **such that the lost,** $L(\hat{m}, \hat{b})$**, is minimal**:

$$(\hat{m}, \hat{b}) = \operatorname{argmin}_{m,b} L(m, b).$$

Finding the optimal values $(\hat{m}, \hat{b})$ is called *fitting the linear model*.

Choosing a Fitness Criterion

Question: What do we mean by "the best line"?

Answer: A line which minimizes the overall error.

But which notion of error should we choose (max absolute deviation, sum/average of absolute deviation or sum/average of squared errors)?

The answer depends on *how*, we believe, the "residual" (difference between observed and predicted values) arise.

Choosing a Fitness Criterion

Our belief: The relationship between *price* (P) and *square footage* (A) is linear

$$P = m \cdot A + b \quad \text{(model for theoretical prices)}$$

But, in real-life, due to unpredictable circumstances observed prices differ from our pricing rule by some *random* amount, Res. This random deviation is called *noise*. So our model for observed housing prices is

$$P = m \cdot A + b + \mathrm{Res} \quad \text{(model for observed prices)}$$

A model that accounts for uncertainty or randomness, where the output (P) is not deterministically dependent on input (A), is called a *statistical model*. The noise, Res, is a *random variable*.

Stats Review

Random Variables

A *random variable* (RV) is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables:

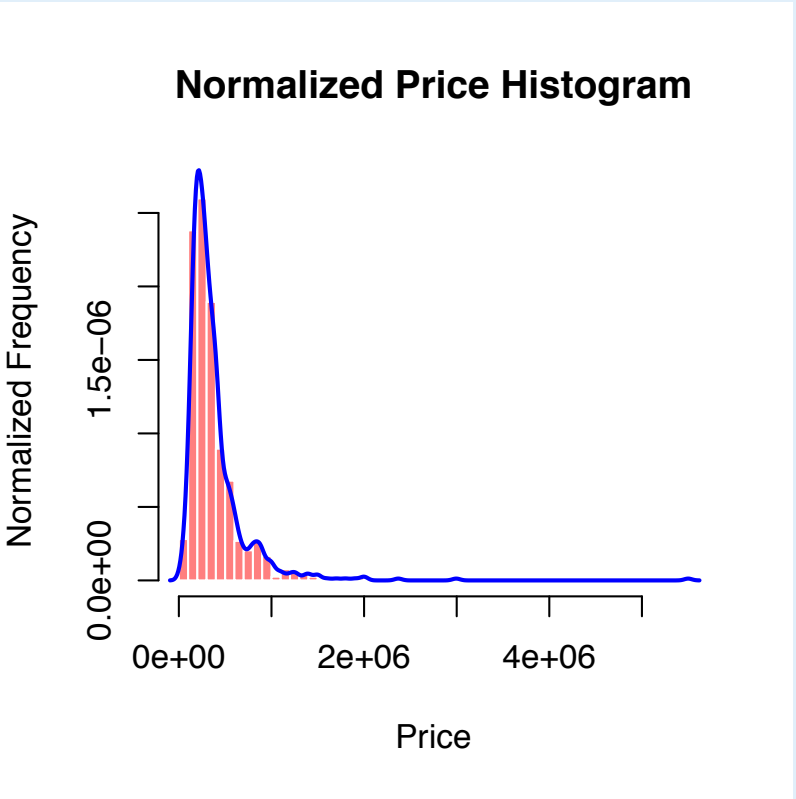
1. a *discrete RV* takes on a finite or countable number of values.

   Ex: The number of bedrooms, B, of a home in our dataset is a random variable. B is discrete.

2. a *continuous RV* usually takes on all values in some range (a, b).

   Ex: The observed price, P, of a home given the square footage is a random variable. P is continuous (can take on all values between 0 and ∞).

Probability Distributions

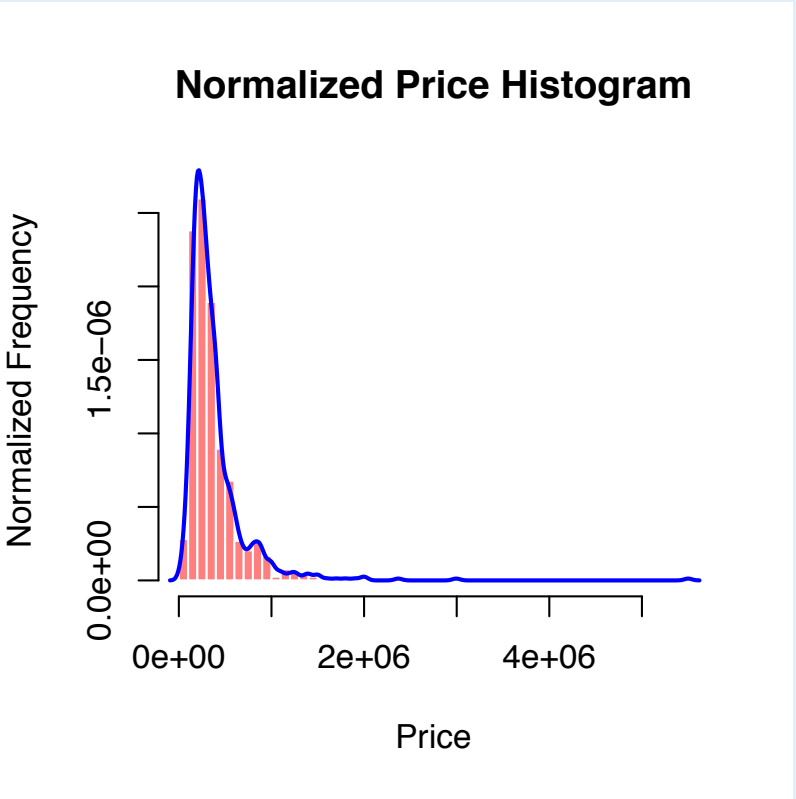


The *probability distribution* of a continuous RV X is given by a function, $p(X)$. The area under p over (a, b) describes the probability of observing values between $X = a$ and $X = b$.

p is called the *probability density function* (pdf) of X.

```
weighted_die = c(0.1, 0.2, 0.3, 0.25, 0.1, 0.05)
barplot(weighted_die, width=1, col=rgb(1,0,0,0.5), xlab="Face",
ylab="Probability", border="white", main="Probabilities of Rolling a Weighted
Die")
```

Probability Distributions

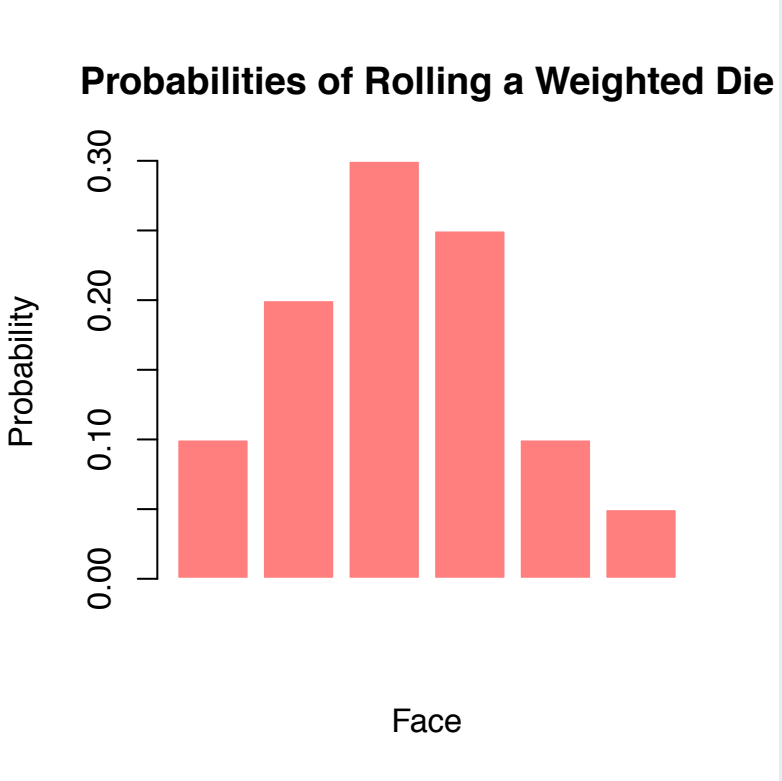


The pdf can provide intuition for how the RV behaves.

For example, the pdf gives us a sense of which values are more likely to be observed compared to others.

```
weighted_die = c(0.1, 0.2, 0.3, 0.25, 0.1, 0.05)
barplot(weighted_die, width=1, col=rgb(1,0,0,0.5), xlab="Face",
ylab="Probability", border="white", main="Probabilities of Rolling a Weighted
Die")
```

Probability Distributions



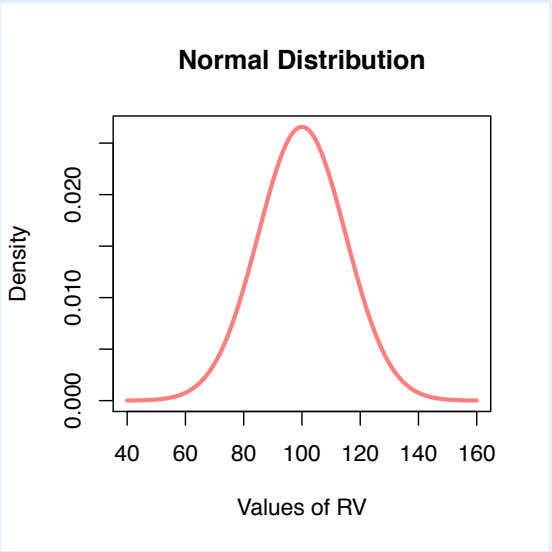
The *probability distribution* of a discrete RV X is given a function, $p(X)$. $p(a)$, written $p(X = a)$, is the probability of observing $X = a$.

p is called the *probability density function* (pdf) or the *probability mass function* (pmf) of X.

```
weighted_die = c(0.1, 0.2, 0.3, 0.25, 0.1, 0.05)
barplot(weighted_die, width=1, col=rgb(1,0,0,0.5), xlab="Face",
ylab="Probability", border="white", main="Probabilities of Rolling a Weighted
Die")
```

Common Types of Distributions (Continuous)

Normal Distribution



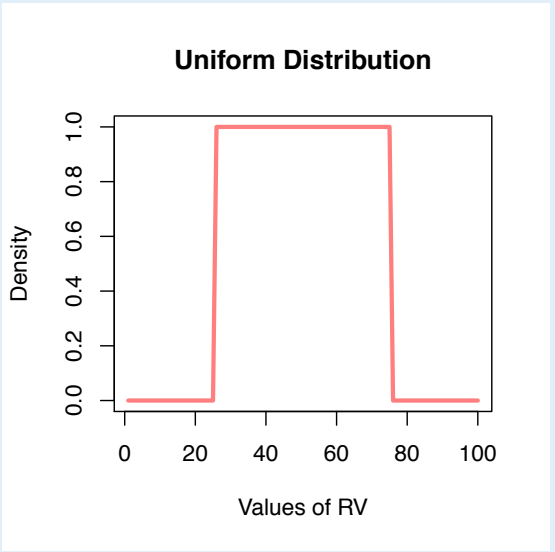
$$p(X) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left\{-\frac{(X-\mu)^2}{2\sigma^2}\right\}$$

$\mu \in \mathbb{R}$, $\sigma > 0$

$X \in (-\infty, \infty)$

$X \sim \mathcal{N}(\mu, \sigma)$

Uniform Distribution



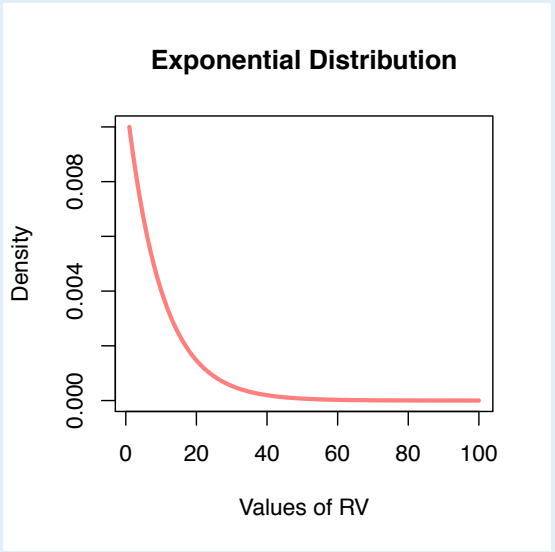
$$p(X) = \begin{cases} \frac{1}{b-a}, & a \leq X \leq b \\ 0, & \text{otherwise} \end{cases}$$

$a, b \in \mathbb{R}$

$X \in [a, b]$

$X \sim \mathrm{U}(a, b)$

Exponential Distribution



$$p(X) = \lambda e^{-\lambda X}$$

$\lambda > 0$

$X \in [0, \infty)$

$X \sim \mathrm{Exp}(\lambda)$

Descriptive Statistics

For many distributions, we can completely describe the shape of the pdf using just a few quantities. These quantities are usually:

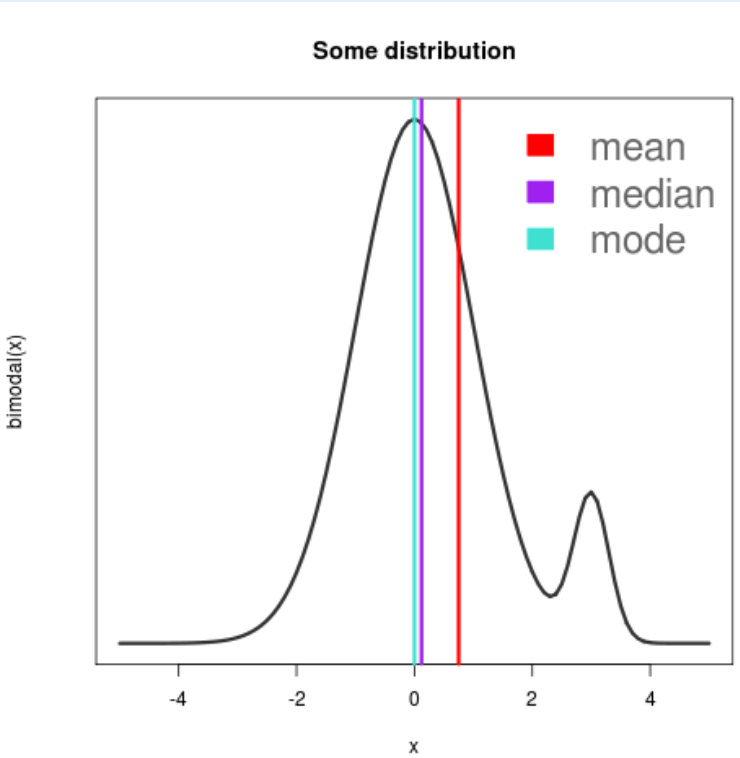
1. (**Measuring the "center"**) The *mean* measures the average of the outcomes, weighted by how likely is the each outcome. The *median* divides the area under the pdf into two equal parts.
2. (**Measuring the "peak"**) The *mode* is the outcome that is the most likely (gives the highest value for the pdf).
3. (**Measuring the "spread"**) The *variance* is measures the average difference between outcomes and the mean, weighted by how likely is each outcome.

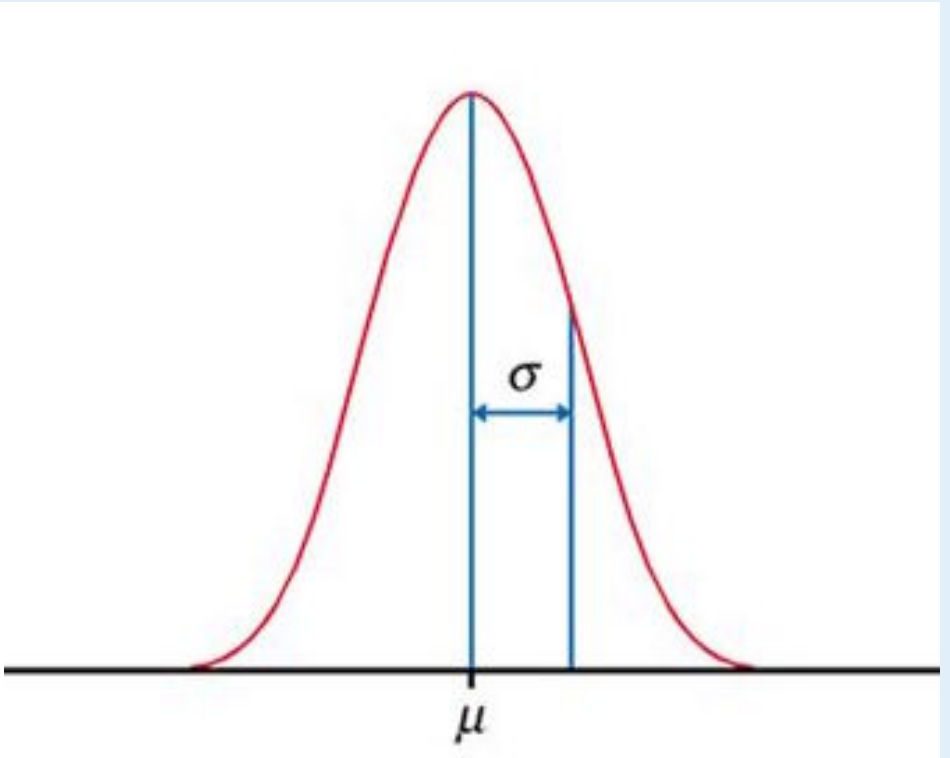
Descriptive Statistics

1. The *mean*, or *expected value*, of a random variable X with some distribution is
$$\mu(X) = \mathbb{E}[X] = \sum_i x_i p(x_i), \int_{-\infty}^{\infty} x p(x) dx$$
2. The *median* is a value $m \in \mathbb{R}$ such that
$$\int_{-\infty}^{m} p(x) dx = \frac{1}{2}$$
3. The *mode* is
$$\mathrm{argmax}_x p(x)$$
4. The *variance* is
$$Var(X) = \mathbb{E}[(X - \mu)^2] = \sum_i (x_i - \mu)^2 p(x_i), \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx$$

Descriptive Statistics

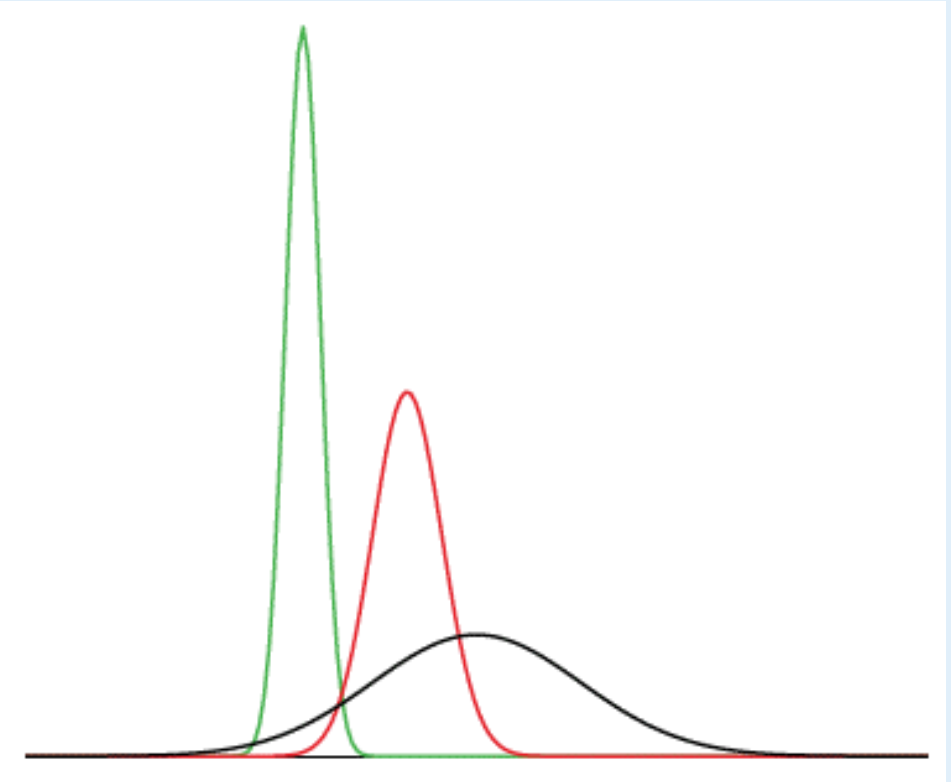
For many distributions, we can completely describe the shape of the pdf using just a few quantities.





Descriptive Statistics

Rank (in increasing order) the following distributions by mean, mode and variance.



New Distributions from Old

Given two random variable X and Y, each with some distribution, we can create new random variables.

1. (**Joint**) The *joint* RV, (X, Y), is a variable that records the values of *both* X and Y, observed simultaneously.
2. (**Conditional**) The *conditional* RV, $X|Y = y$ (or simply $X|Y$), is a variable that records the observed value of X *after* observing that $Y = y$.

New Distributions from Old

Given two random variable X and Y, each with some distribution, we can create new random variables. We can also compute the distribution of these new RVs.

1. $p(X, Y) = p(X|Y)p(Y) = p(Y|X)p(X)$
   (what happens if X and Y are independent?)
2. $p(X|Y) = \dfrac{p(X, Y)}{p(Y)}$
3. $p(Y|X) = \dfrac{p(X|Y)p(Y)}{p(X)}$, (Baye's Theorem)

An Example

Recall that our model for observed housing prices, P, given square footage, A is

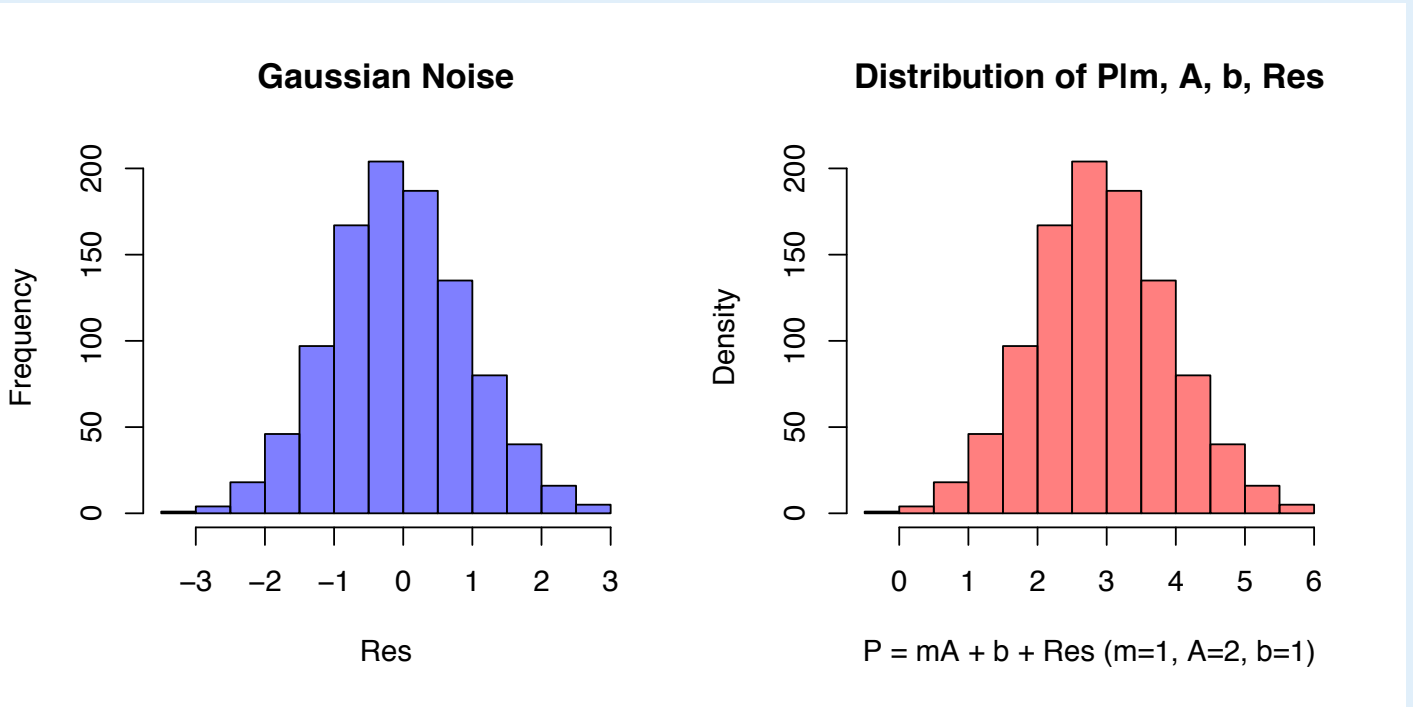
$$P = m \cdot A + b + \mathrm{Res} \quad \text{(model for observed prices)}$$

where Res is a random noise variable.

What kind of distributions involving the RV P can we derive from our model?

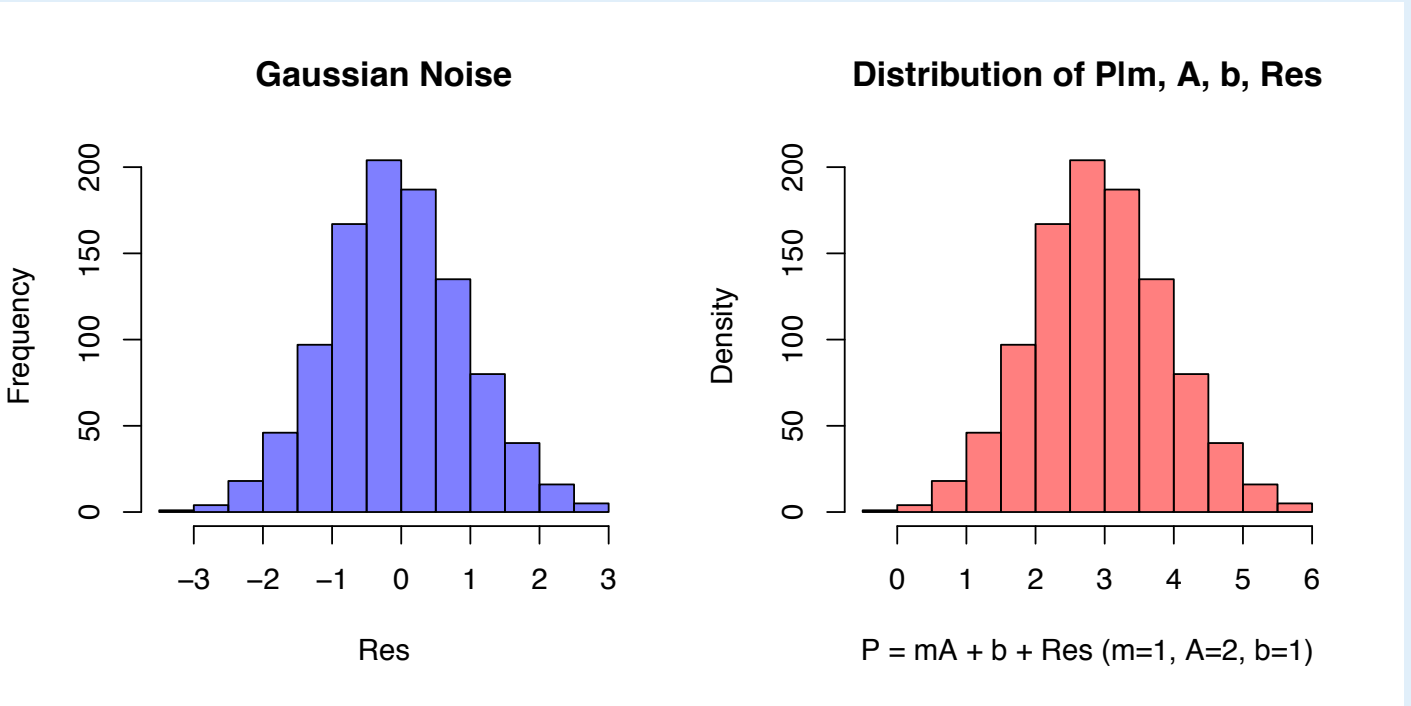
An Example

Let's fix m, b, A and choose a normal distribution for Res, say $\mathrm{Res} \sim \mathcal{N}(0, 1)$. What is the distribution of $P|m, b, A, \mathrm{Res}$?



An Example

Let's fix m, b, A and choose a normal distribution for Res, say $\mathrm{Res} \sim \mathcal{N}(0, 1)$. What is the distribution of $P|m, b, A, \mathrm{Res}$?



$$P|m, b, A, \mathrm{Res} \sim \mathcal{N}(m \cdot A + b, 1)$$

Linear Regression (Univariate)

Back to Our Linear Model

Recall that our statistical model for observed housing prices is

$$P = m \cdot A + b + \mathrm{Res}$$

Suppose that $\mathrm{Res} \sim \mathcal{N}(0, 1)$, then $P|m, b, A, \mathrm{Res} \sim \mathcal{N}(m \cdot A + b, 1)$. Let's say we have observed a single home listing $A = 1000$ and $P = \$1.0$ mil.

What does it mean to *fit our model to the data*?

Hint: Use `R` to calculate $p(P|m, b, A, \mathrm{Res})$ for a couple of choices of m and b.

Back to Our Linear Model

Recall that our statistical model for observed housing prices is

$$P = m \cdot A + b + \mathrm{Res}$$

Suppose that $\mathrm{Res} \sim \mathcal{N}(0, 1)$, then $P|m, b, A, \mathrm{Res} \sim \mathcal{N}(m \cdot A + b, 1)$. Let's say we have observed a single home listing $A = 1000$ and $P = \$1.0$ mil.

What does it mean to *fit our model to the data*?

Let's find m_{MLE} and b_{MLE} so that $p(P|m_{MLE}, b_{MLE}, A, \mathrm{Res})$ is maximal. I.e. the model $P = m_{MLE} \cdot A + b_{MLE} + \mathrm{Res}$ explains the observed data with the highest probability.

The above model is called the *maximum likelihood estimator (MLE)*.

Ordinary Least Squares

For Gaussian noise, $\mathrm{Res} \sim \mathcal{N}(0,1)$, finding the MLE model is exactly the same as minimizing the sum of squared residuals!

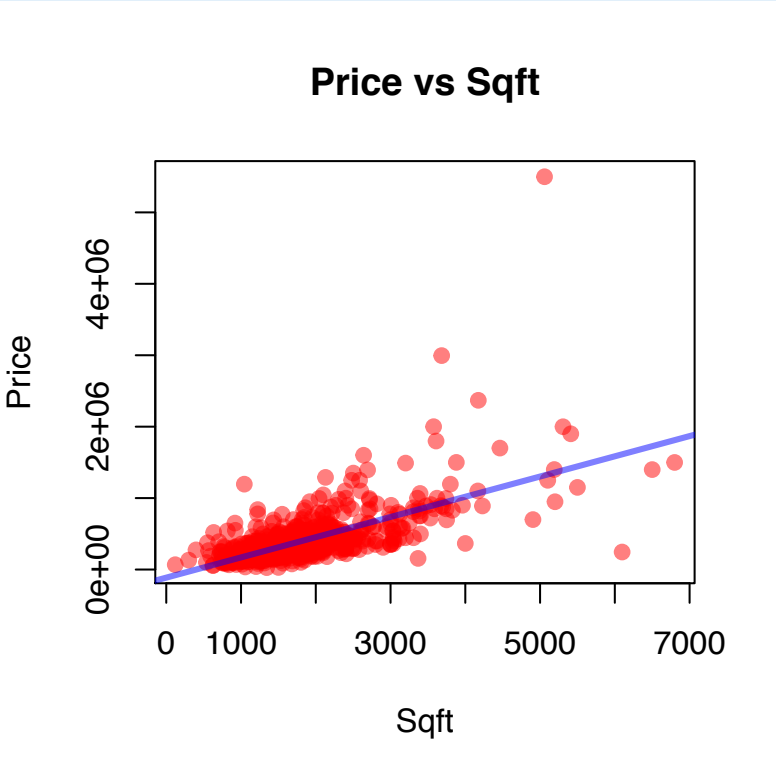
Given a set of points $(A_1, P_1), \ldots, (A_n, P_n)$, the sum of squared deviations of a linear model $P = mA + b$ is

$$L(m,b) = \sum_{i=1}^{n} |P_i - (mA_i + b)|^2$$

L is called the lost (or cost) function. Our goal is to **find m_{MLE} and b_{MLE} such that the lost, $L(m_{MLE}, b_{MLE})$, is minimal**:

$$(m_{MLE}, b_{MLE}) = \mathrm{argmin}_{m,b} L(m,b).$$

Linear Regression in R




```
mydata.lm <- lm(Price~Sqft,
data=mydata)
coeffs <- coefficients(mydata.lm);
coeffs
plot(mydata$Sqft, mydata$Price,
main="Price vs Sqft",
col=rgb(1,0,0,0.5), xlab="Sqft",
ylab="Price", pch=19)
abline(mydata.lm,
col=rgb(0,0,1,0.5), lwd=3)
sm <- summary(mydata.lm)
SSR <- mean(sm$residuals^2); SSR
```

Question: The regression line is the "best-fitting" linear model to the data. But just how good is it?

Evaluating Predictors

Given a set of home listings, we can now fit a maximum likelihood linear model

$$P = m_{MLE} \cdot A + b_{MLE}$$

by minimizing the sum of squared residuals (ordinary least squares or OLS). The "error" made by our model in fitting the data is

$$L(m_{MLE}, b_{MLE}) = \sum_{i=1}^{n} |P_i - (m_{MLE} \cdot A_i + b_{MLE})|^2 .$$

This is called the *training error*.

But we also need to evaluate our model on new data that it has not yet seen, *test data*.

Evaluating Predictors

1. Given a set of data $(A_1, P_1), \ldots, (A_n, P_n)$, split the data into a *training set* and a *test set*.

```
train.size <- floor(0.70 * nrow(mydata))
set.seed(100)
train.ind <- sample(seq_len(nrow(mydata)), size = train.size)
train <- mydata[train.ind, ]; test <- mydata[-train.ind, ]
```

2. Fit the model on the training set, report error

```
train.lm <- lm(Price~Sqft, data=train)
sm <- summary(train.lm)
SSR.train <- mean(sm$residuals^2); SSR.train
```

3. Fit the model on the testing set, report error

```
pred <- predict(train.lm, test)
SSR.test <- mean((pred - test$Price)^2); SSR.test
```

Linear Regression (Multivariate)

Linear Regression in Multiple Variables

It's a little bit unreasonable for price of a home to depend on square footage alone. In reality, P most likely depends on some combination of square footage, A, number of bedrooms Bd and the number of bathrooms Ba.

The easiest relationship between all 4 variables is again linear

$$P = a_0 + a_1 A + a_2 Bd + a_3 Ba + \mathrm{Res}.$$

Again, if we take $\mathrm{Res} \sim \mathcal{N}(0, 1)$, P is a random variable with a normal distribution, $P \sim \mathcal{N}(a_0 + a_1 A + a_2 Bd + a_3 Ba, 1)$.

Just like before, the values of $a_0, \ldots, a_3$ which maximizes the likelihood of the data (MLE model) can be found by minimizing sum of squared residuals.

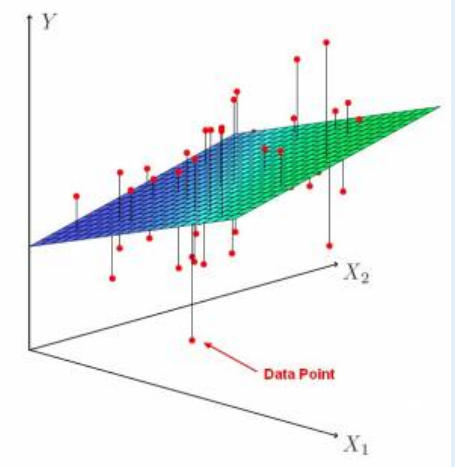
Multiple Linear Regression in R

Multiple linear regression for our model:

```
train.lm <- lm(Price~Sqft + Bedrooms + Bathrooms, data=train)
```

In general, for variables $y, x_1, \ldots, x_n$ in your data frame `train`:

$$\texttt{train.lm} \leftarrow \texttt{lm}(y \sim x_1 + \ldots + x_n, \texttt{data} = \texttt{train})$$



Polynomial Regression

Polynomial Regression

As we've noticed, our linear models (univariate and multivariate) don't seem to fit the housing data very well.

Maybe this is because the underlying relationship between price and square footage (or number of rooms) isn't linear. Perhaps the model we want is polynomial

$$P = a_0 + a_1 A + a_2 A^2 + \mathrm{Res}, \quad \mathrm{Res} \sim \mathcal{N}(0, 1)$$

Fitting a polynomial model (of degree d) in `R`

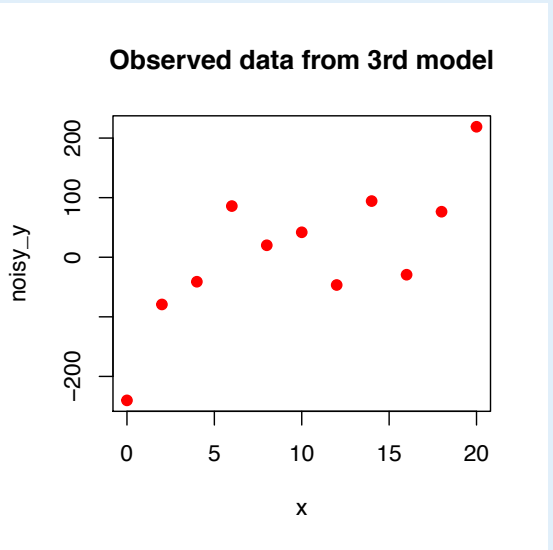
```
train.pm <- lm(Price~poly(Sqft, d), data=train)
```

Exercise: Try to fit the model with a few values of d and report the training and testing errors.

Overfitting

One thing we notice immediately was that the error on the training set decreases as the degree of the polynomial increases.

So maybe we generally want to pick very high degree polynomials to model our data?

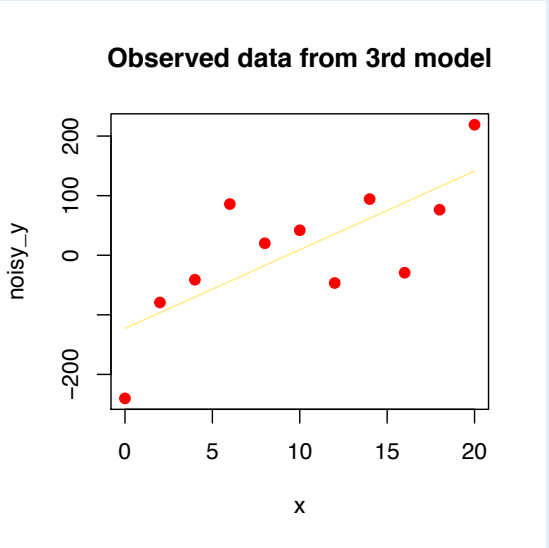


What is happening to our model as the degree increases?

Overfitting

One thing we notice immediately was that the error on the training set decreases as the degree of the polynomial increases.

So maybe we generally want to pick very high degree polynomials to model our data?

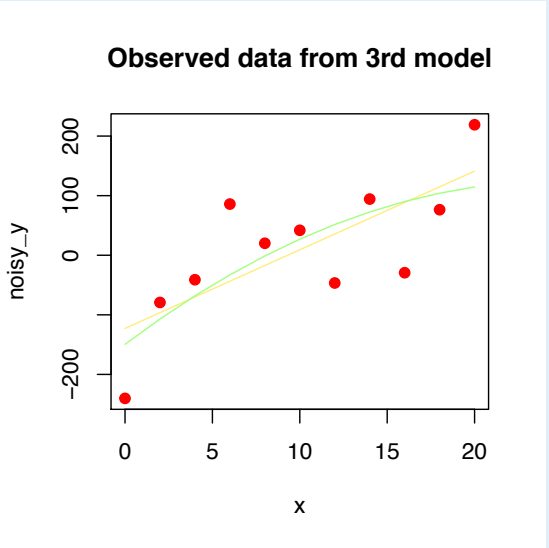


What is happening to our model as the degree increases?

Overfitting

One thing we notice immediately was that the error on the training set decreases as the degree of the polynomial increases.

So maybe we generally want to pick very high degree polynomials to model our data?

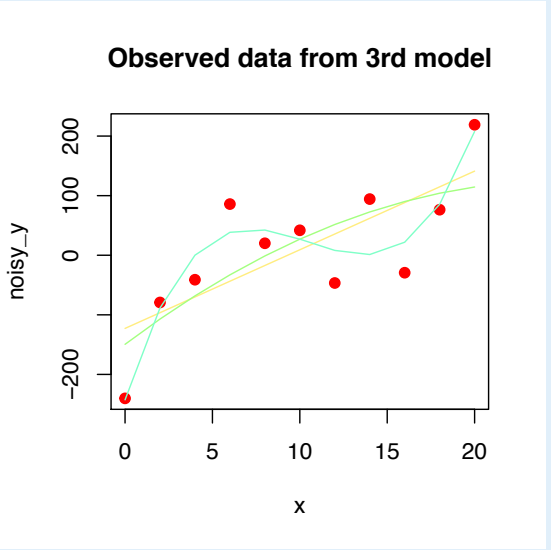


What is happening to our model as the degree increases?

Overfitting

One thing we notice immediately was that the error on the training set decreases as the degree of the polynomial increases.

So maybe we generally want to pick very high degree polynomials to model our data?

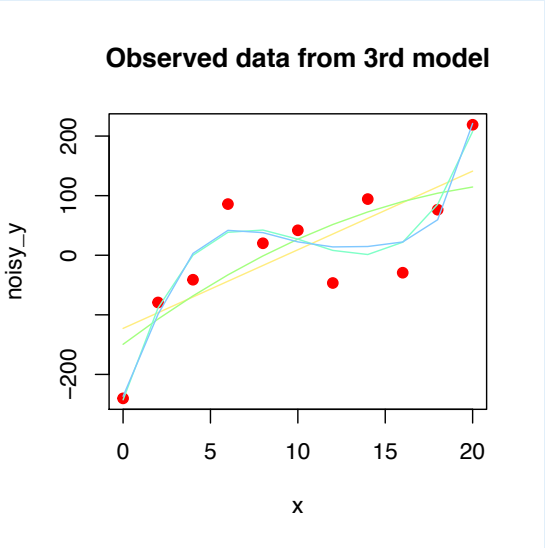


What is happening to our model as the degree increases?

Overfitting

One thing we notice immediately was that the error on the training set decreases as the degree of the polynomial increases.

So maybe we generally want to pick very high degree polynomials to model our data?

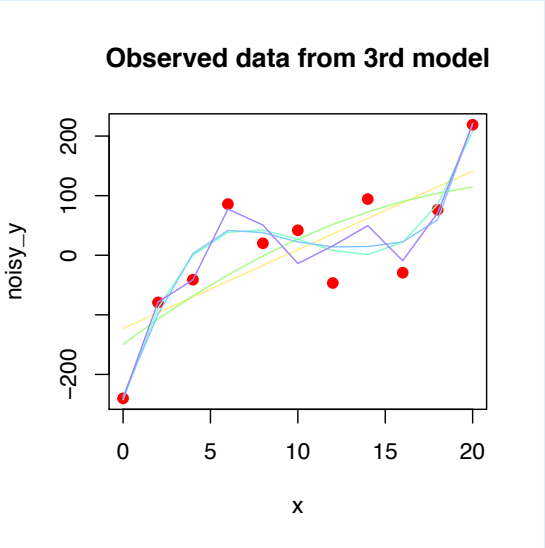


What is happening to our model as the degree increases?

Overfitting

One thing we notice immediately was that the error on the training set decreases as the degree of the polynomial increases.

So maybe we generally want to pick very high degree polynomials to model our data?

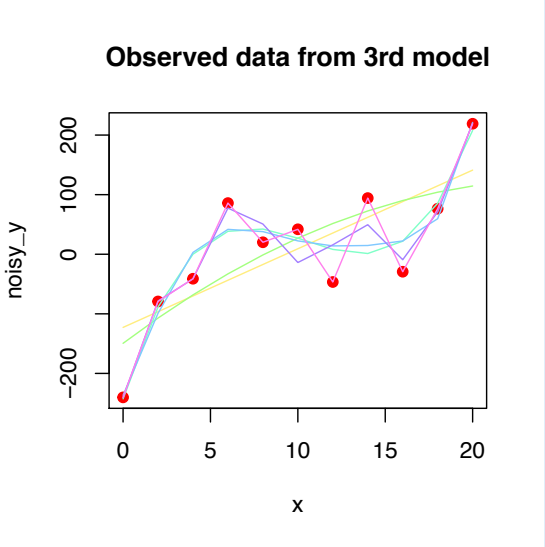


What is happening to our model as the degree increases?

Overfitting

One thing we notice immediately was that the error on the training set decreases as the degree of the polynomial increases.

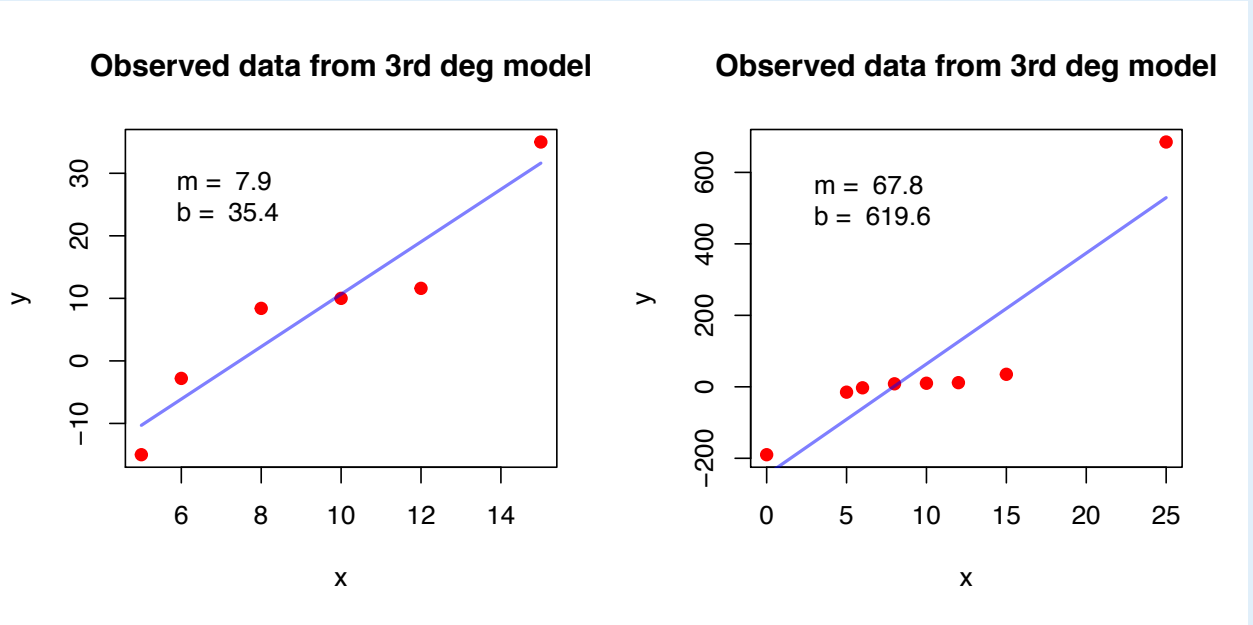
So maybe we generally want to pick very high degree polynomials to model our data?



What is happening to our model as the degree increases?

Overfitting

Overfitting can happen with linear regression too!



What happens to our linear model as we add two new data points?

In multiple linear regression, what happens when we have N number of observations and N number of explanatory variables?

Overfitting

Overfitting happens when we learn parameters or rules that are too specific to the training set, so much that our model is not useful in explaining new data (we do great on train data but poorly on test).

Overfitting can happen when we have too few observations compared to the number of variables in our model with which we try to explain the observations.

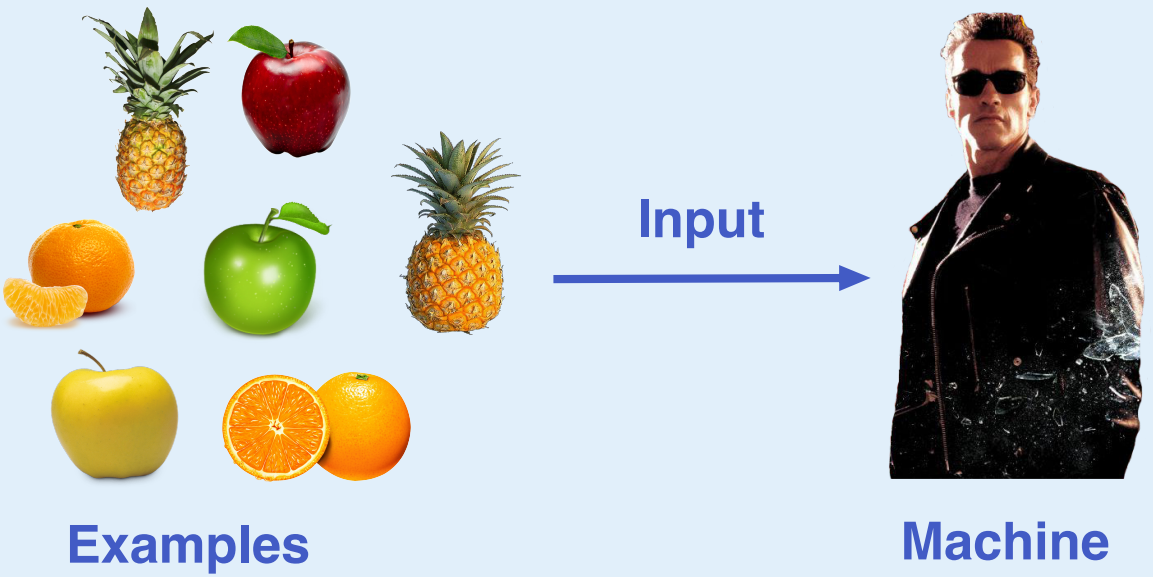
Later, we'll see that overfitting can be curbed by regularization and variable selection.

What is Machine Learning

Intuition

The goal of machine learning is to be able to teach machines to make decisions based on previous experience, just like humans.

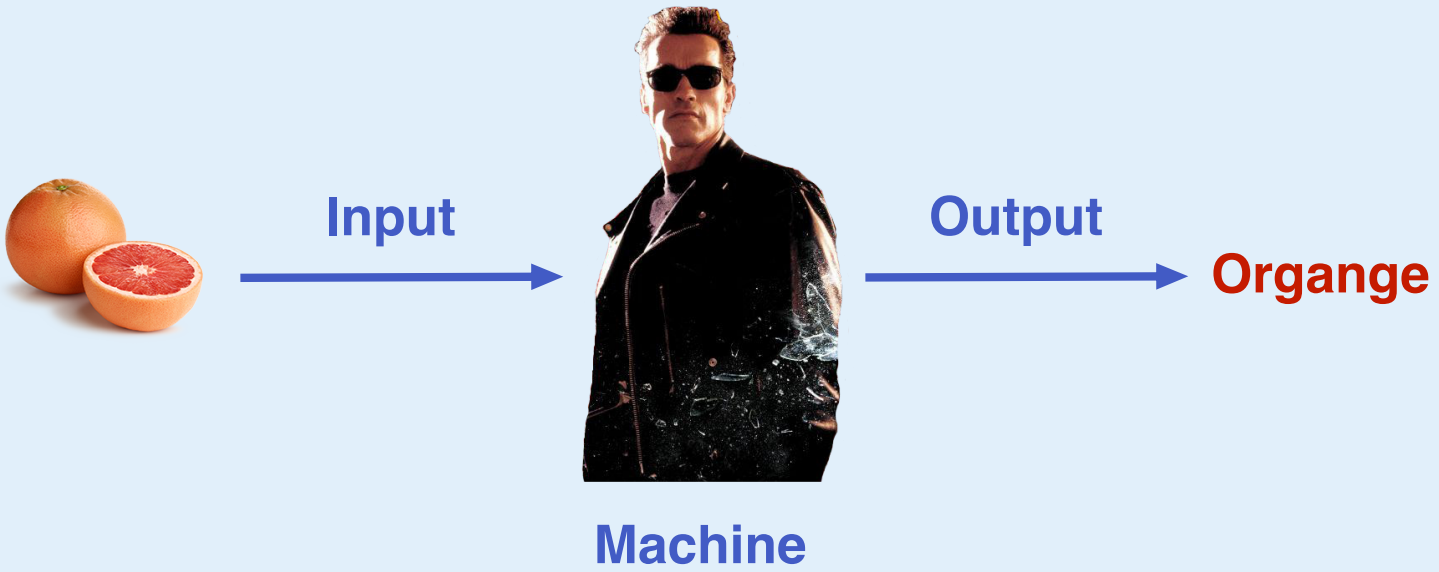
For example, we give the machine examples of a type of object or scenario.



Intuition

The goal of machine learning is to be able to teach machines to make decisions based on previous experience, just like humans.

For example, we give the machine examples of a type of object or scenario. We hope that machine will recognize a new instance of that type of object or scenario.

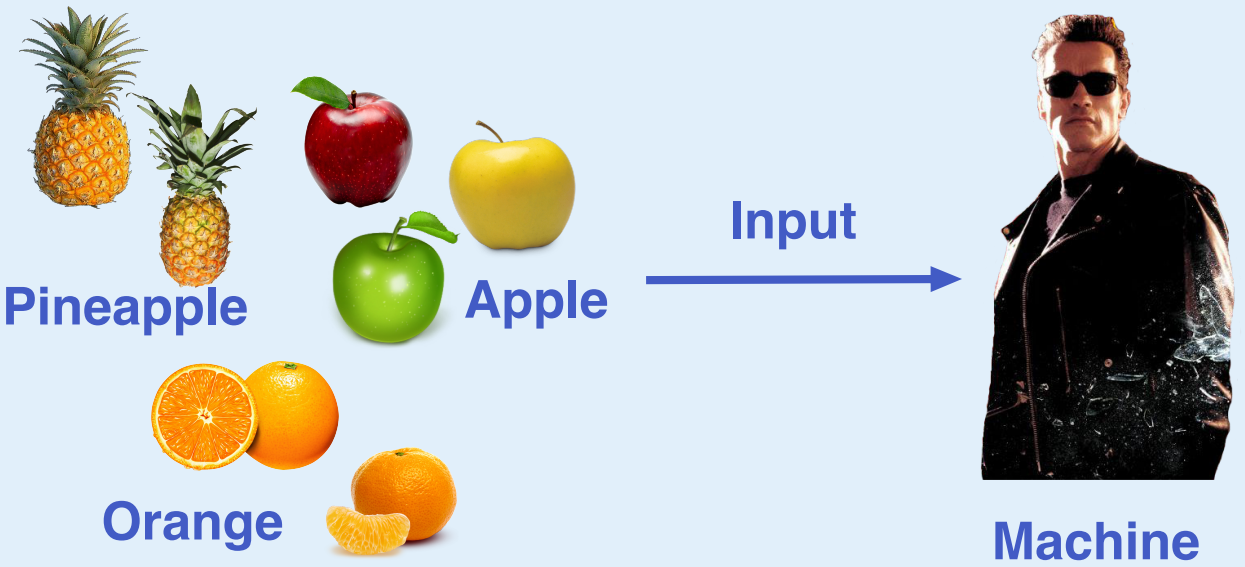


Intuition

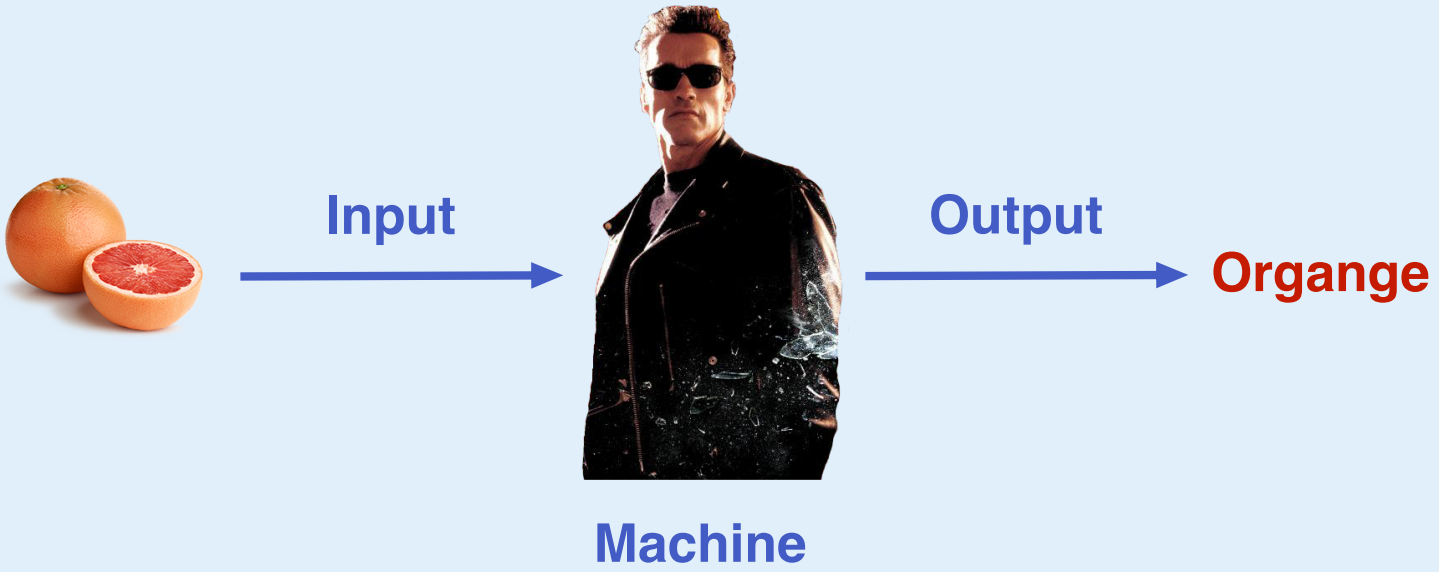
The goal of machine learning is to be able to teach machines to make decisions based on previous experience, just like humans.

The "learning" done by the machine is usually fitting a statistical model to a set of training data. The machine can then use this calibrated model to make decisions when encountering new data.

Supervised Learning



Supervised Learning



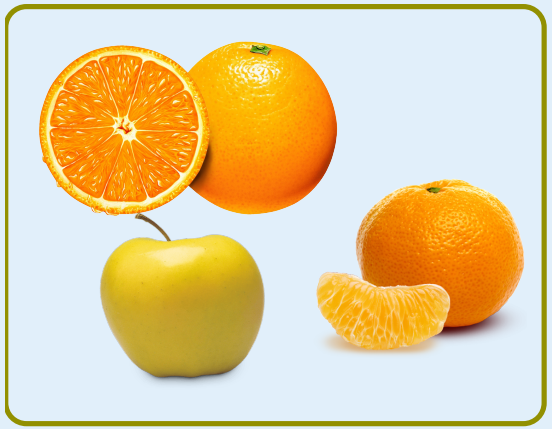
Unsupervised Learning

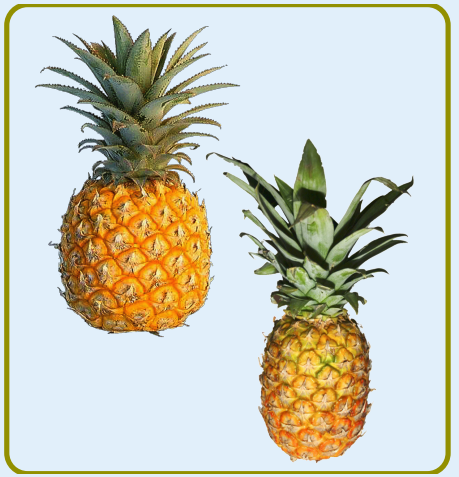
Examples

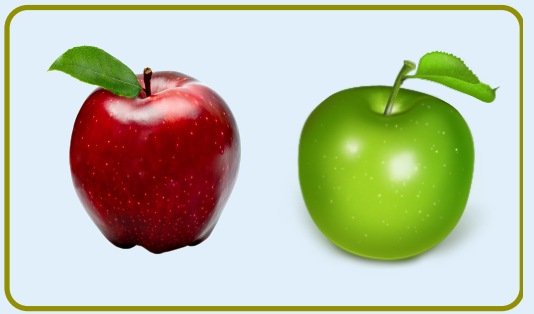
Input

Machine

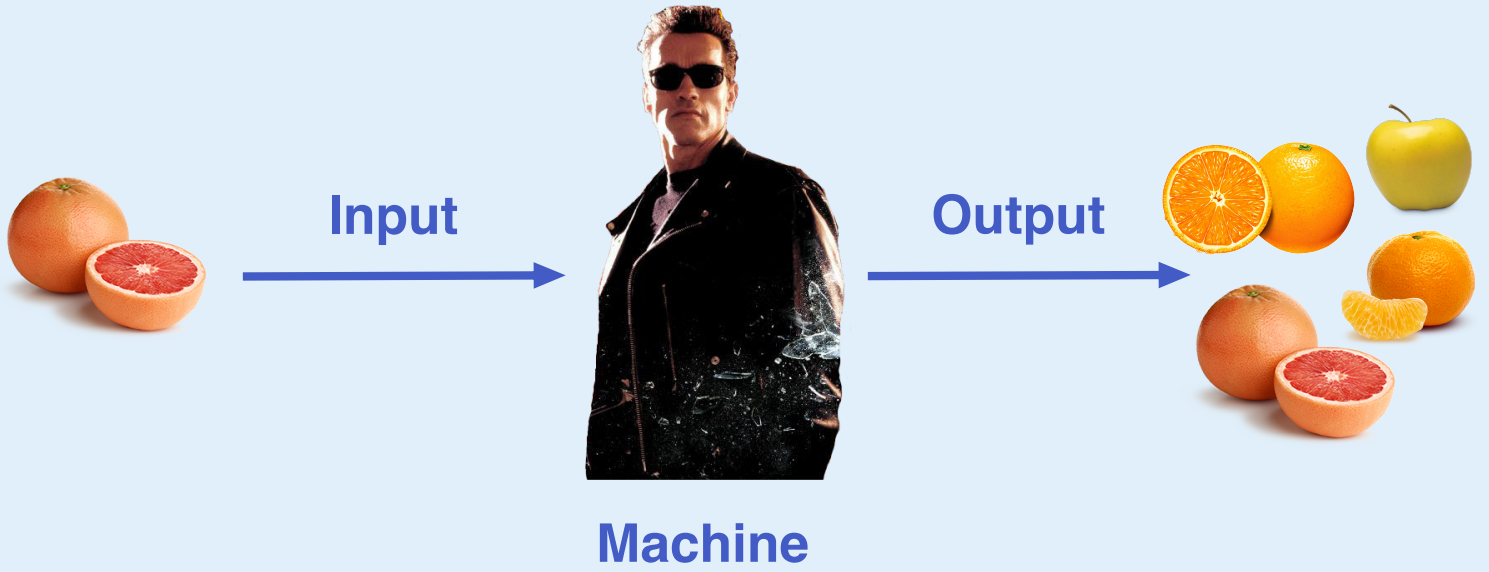
Unsupervised Learning







Unsupervised Learning



Final Thoughts

Question: Is regression (linear or polynomial) supervised or unsupervised learning? Why?

Question: What might be the pros and cons of supervised (resp unsupervised learning)? Why?

Question: Why is it important to use statistical models in representing data?