March 3: Data, models, errors
Questions for today

- How can we \textit{filter} a pandas data frame?

- Why are squared errors important, and how do they relate to the normal distribution and log likelihood?

- How can we predict one variable given another? \textit{What makes avocados cost more or less?}

- How do we compare predictive models?
Questions for today

• How can we filter a pandas data frame?

• Why are squared errors important, and how do they relate to the normal distribution and log likelihood?

• How can we predict one variable given another?
  What makes avocados cost more or less?

• How do we compare predictive models?
Normal / Gaussian Probability

\[ p(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x-\mu)^2}{2\sigma^2} \right] \]
Normal / Gaussian Probability

\[
p(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)
\]
Normal / Gaussian Probability

\[ p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \]
Normal/Gaussian Probability

\[ P(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \]

↑
constant that makes the rest add up to 1.0
Normal log likelihood

\[
\log p(x_1, x_2, x_3, \ldots, x_n|\mu, \sigma^2) = \log \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x_i-\mu)^2}{2\sigma^2} \right] = N \cdot \left( -\frac{1}{2} \log(2\pi\sigma^2) \right) - \sum_i \frac{(x_i-\mu)^2}{2\sigma^2}
\]
Normal log likelihood

\[
\log p(x_1, x_2, x_3, \ldots, x_N | \mu, \sigma^2) \\
= \log \prod_{i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x_i - \mu)^2}{2\sigma^2} \right] \\
= N \cdot \left( -\frac{1}{2} \log(2\pi\sigma^2) \right) - \sum_{i} \frac{(x_i - \mu)^2}{2\sigma^2}
\]

\( \prod \) and \( \sum \) are for-loops
Normal log likelihood

\[
\log p(x_1, x_2, x_3, \ldots, x_n | \mu, \sigma^2) = \log \prod \left[ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right]
= N \cdot \left( -\frac{1}{2} \log(2\pi\sigma^2) \right) - \sum \frac{(x_i - \mu)^2}{2\sigma^2}
\]
Normal log likelihood

\[ \log p(x_1, x_2, x_3, ..., x_n | \mu, \sigma^2) \]

\[ = \log \prod_i \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left[ -\frac{(x_i - \mu)^2}{2\sigma^2} \right] \]

\[ = N \cdot \left( -\frac{1}{2} \log(2\pi \sigma^2) \right) - \sum_i \frac{(x_i - \mu)^2}{2\sigma^2} \]

---

'\text{doesn't involve}\ x_i \text{ or } \mu'

'SUM OF SQUARES!!'
\[- \sum_{i} \frac{(x_i - \mu)^2}{2\sigma^2}\]

Log likelihood increases when squared distance to mean decreases.
Questions for today

- How can we filter a pandas data frame?
- Why are squared errors important, and how do they relate to the normal distribution and log likelihood?
- How can we predict one variable given another? What makes avocados cost more or less?
- How do we compare predictive models?
data = model + error
data = model + error

y_i = \mu + \epsilon_i
data = model + error

\[ y_i = \mu + \epsilon_i \]

Each \( i \)th data point has its own error.
data = model + error

\[ y_i = \mu_i + \epsilon_i \]

all observations share the same mean
What will avocados cost?

Los Angeles

Syracuse

$\$\$\$?
Model 0:

\[ \text{price}_i = \mu_i + \varepsilon_i \]

Model 1:

\[ \text{price}_i = M_{\text{city}_i} + \varepsilon_i \]
Model 0:
\[ \text{price}_i = \mu + \varepsilon_i \]
both cities have same mean price

Model 1:
\[ \text{price}_i = \mu_{\text{city}_i} + \varepsilon_i \]
each city has its own mean price
if $M_{SYR} \neq M_{LA}$,

What will a histogram of prices look like?
Compare models using squared error

$$\sum_i (\text{price}_i - \text{prediction}_i)^2$$

Which model will have less error?