Introduction to Data Science

Introduction to Modeling and Algorithms

Joanna Bieri DATA101

Important Information

- Email: joanna_bieri@redlands.edu
- Office Hours: Duke 209 Click Here for Joanna's Schedule

Announcements

If you are behind it is time to get 100% caught up! I will expect to see you at one of the labs!

Day 15 Assignment - same drill.

1. Make sure you can **Fork** and **Clone** the Day15 repo from Redlands-DATA101
2. Open the file Day15-HW.ipynb and start doing the problems.
   - You can do these problems as you follow along with the lecture notes and video.
3. Get as far as you can before class.
4. Submit what you have so far **Commit** and **Push** to Git.
5. Take the daily check in quiz on **Canvas**.
6. Come to class with lots of questions!

What is a Mathematical Model?

A model is used to explain the relationship between variables so that we can make predictions.

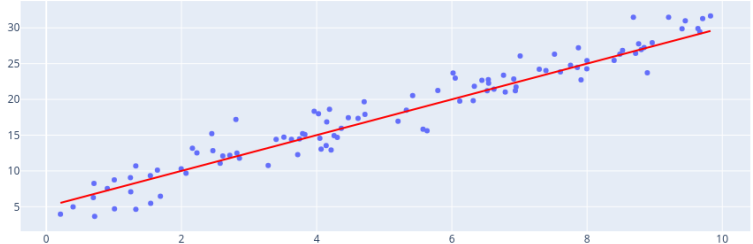
- Is there a relationship between A and B?
- If I knew A could I predict B?

This lab follows the Data Science in a Box units "The Language of Models and Fitting and Interpreting Models" by Mine Çetinkaya-Rundel. It has been updated for our class and translated to Python by Joanna Bieri.

Linear Models:

A **linear model** can be described by a straight line.

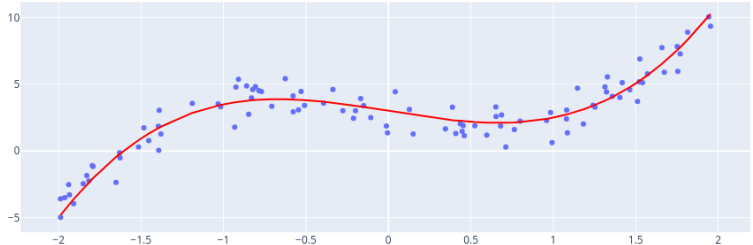
$$Y = mX + b$$



Nonlinear Models:

A **nonlinear model** cannot be described by a straight line. It might be modeled by a wide range of other functions!

$$Y = f(X)$$



Paris Paintings Data

- Source: Printed catalogs of 28 auction sales in Paris, 1764 - 1780 (Historical Data)
- Data curators Sandra van Ginhoven and Hilary Coe Cronheim (who were PhD students in the Duke Art, Law, and Markets Initiative at the time of putting together this dataset) translated and tabulated the catalogs
- 3393 paintings, their prices, and descriptive details from sales catalogs over 60 variables

Historical Example

Two paintings very rich in composition, of a beautiful execution, and whose merit is very remarkable, each 17 inches 3 lines high, 23 inches wide; the first, painted on wood, comes from the Cabinet of Madame la Comtesse de Verrue; it represents a departure for the hunt: it shows in the front a child on a white horse, a man who gives the horn to gather the dogs, a falconer and other figures nicely distributed across the width of the painting; two horses drinking from a fountain; on the right in the corner a lovely country house topped by a terrace, on which people are at the table, others who play instruments; trees and fabriques pleasantly enrich the background.

Historical Example

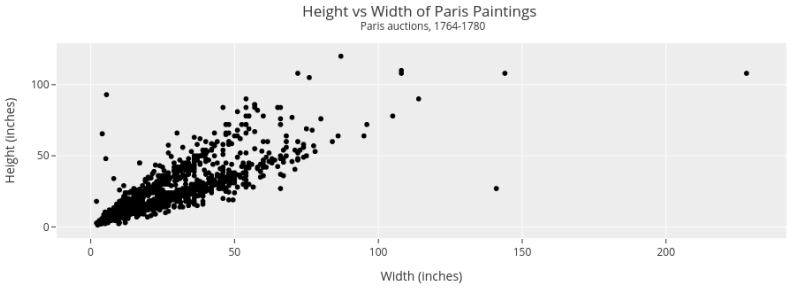
Lets look at the information for the painting described above.

	name	sale	lot	position	dealer	year	origin_author
2518	R1777-89a	R1777	89	0.375527	R	1777	D/FL

Models as Functions

- We can represent relationships between variables using **functions**
- A function is a mathematical concept: the relationship between an output and one or more inputs

Can we find a function that describes the relationship between height and width of painting?



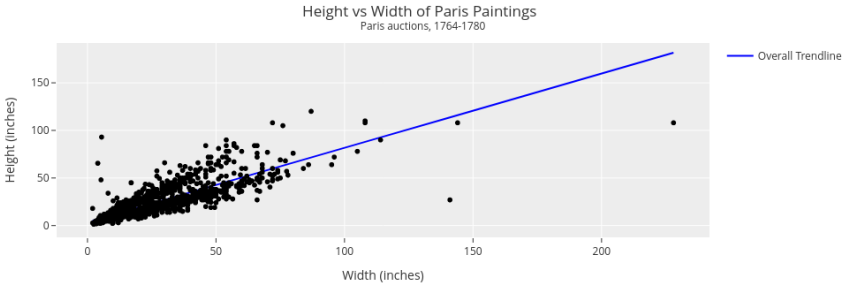
Add a trendline

`trendline='ols'`

This uses Ordinary Least Squares fitting to find a reasonable line.

So the line that "fits" this data based on the code we ran is

$$H = 0.7808W + 3.6214$$



Did we have to pick this trendline?

No there are other choices:

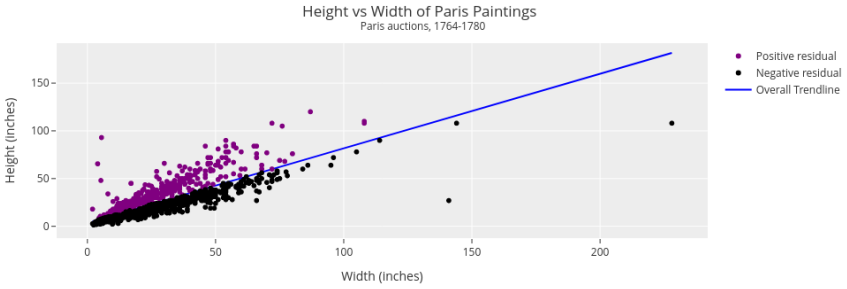
- 'ols': Ordinary Least Squares regression line (linear regression).
- 'lowess': Locally Weighted Scatterplot Smoothing (non-linear regression).
- 'rolling': Rolling window calculations (e.g., rolling average, rolling median).
- 'expanding': Expanding window calculations (e.g., expanding average, expanding sum).
- 'ewm': Exponentially Weighted Moving Average (EWMA).

Some of these we will cover in the next few weeks, but some will wait for DATA 201.

Vocab for Modeling

- **Response variable:** Variable whose behavior or variation you are trying to understand, on the y-axis
- **Explanatory variables:** Other variables that you want to use to explain the variation in the response, on the x-axis
- **Predicted value:** Output of the **model function**
  - The model function gives the typical (expected) value of the response variable *conditioning* on the explanatory variables
- **Residuals:** A measure of how far each case is from its predicted value (based on a particular model)
  - Residual = Observed value - Predicted value
  - Tells how far above/below the expected value each case is

Residual Plot



Models - upsides and downsides

- Models can sometimes reveal patterns that are not evident in a graph of the data. This is a great advantage of modeling over simple visual inspection of data.
- There is a real risk, however, that a model is imposing structure that is not really there on the scatter of data, just as people imagine animal shapes in the stars. A skeptical approach is always warranted.
- Just because you fit a model does not mean you found a true relationship.

Variation around the model.

Variation or Uncertainty is just as important as the model, if not more!

Statistics is the explanation of variation in the context of what remains unexplained.

- The scatter suggests that there might be other factors that account for large parts of painting-to-painting variability, or perhaps just that randomness plays a big role.
- Adding more explanatory variables to a model can sometimes usefully reduce the size of the scatter around the model.

How do we use models?

TWO MAIN USES:

- Explanation: Characterize the relationship between y and x via *slopes* for numerical explanatory variables or *differences* for categorical explanatory variables
- Prediction: Plug in x, get the predicted y

Linear (Least Squares) Regression - Supervised Machine Learning

Find a straight line that is the best fit for our data:

$$\hat{y}_i = b_0 + b_1 x_i$$

We want to minimize the cost (reduce the sum of the residuals)

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

where y_i is the value from our sample data and $\hat{y}_i$ is the value predicted from our model.

Either exactly or using gradient descent we learn the values of b_0 and b_1 that are best.

Train a Linear Regression Model

This happens in three steps

1. Data Preprocessing
2. Model Training
3. Analyze the output - predictions

```
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import OneHotEncoder
```

Preprocessing the Data

1. Select the variables that you wan to use (columns)
2. Decide what to do about NaNs or other strange data
3. (*advanced*) Think about rescaling and standardizing
4. Create the inputs and outputs (sometimes encode)
5. (*advanced*) Test - Train split

Train the model

1. Create the base model, in this case LinearRegression()
2. Train the model using the training data
3. Look at the results.

```
X = DF_model['Width_in'].values.reshape(-1, 1)
y = DF_model['Height_in'].values

LM = LinearRegression()
LM.fit(X, y)

LM.coef_
LM.intercept_
```

Our Paintings Example

This means that linear regression found the line

$$height_i = 3.6214055418381896 + 0.78079641 * width_i$$

so given a width we could predict a height. There appears to be a positive relationship between height and width - in general as the width increases the height increases.

Now we can use our model to make a prediction

Let's say we know a painting had certain width, we can predict the height

```
width = 33
width = np.array(width).reshape(-1,1)
LM.predict(width)
```

array([29.38768703])

```
width = 0
width = np.array(width).reshape(-1,1)
LM.predict(width)
```

array([3.62140554])

Important values

- **Slope:** For each additional inch the painting is wider, the height is expected to be higher, on average, by 0.781 inches.
- **Intercept:** Paintings that are 0 inches wide are expected to be 3.62 inches high, on average. (Does this make sense?)

Correlation does not imply causation

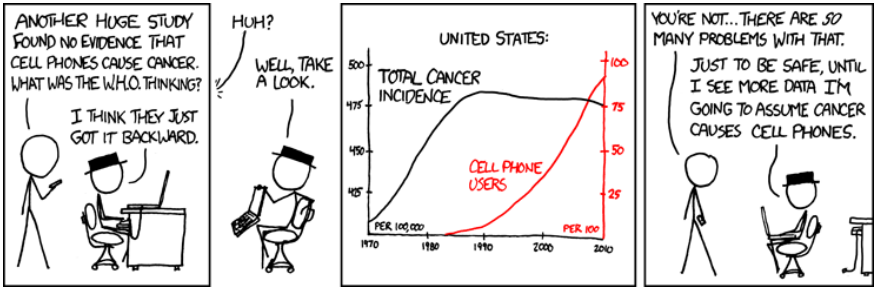


Figure 1: Correlation

Source: XKCD, Cell phones

Check the accuracy of your model

LM.score() uses the R^2 score:

- R^2 is the coefficient of determination
- RSS is the sum of squares of residuals
- TSS is the total sum of squares

We got a score of 0.6829467672722757

What about Categorical Data?

Does the price of a painting depend on whether or not it was landscape?

x = "landscape" or "not landscape"

y = height of painting.

The equation

$$\widehat{y} = b_0 + b_1 \text{"}landscape\text{"}$$

does not make sense.

What about Categorical Data?

0 = "not landscape" 1 = "landscape"

Then we can plug in zero or one into the equation:

$$\hat{y} = b_0 + b_1 landscape$$

What about Categorical Data?

Here is the result of training a linear regression to predict height given whether or not a painting is landscape.

How do we interpret this result?

$$Height = 22.68 - 5.65(landsALL)$$

So if landsALL $= 0$ (not a landscape) the paintings have an average height of 22.68. If landsALL $= 1$ (it is a landscape) then the height on average is reduced by 5.65 inches. We found that landscapes tend to be shorter.

More than two categories - One Hot Encoding

In the case above we could easily code the categorical to just be 0 or 1. Because there were only two options.

If there are more than two options we have to choose a way to encode the data. **one-hot encoding**

More than two categories - School of Painting

(A = Austrian, D/FL = Dutch/Flemish, F = French, G = German, I = Italian, S = Spanish, X = Unknown)

```
my_columns = ['school_pntg','price']
DF_model = DF[my_columns]

X = DF_model['school_pntg'].values.reshape(-1,1)
y = DF_model['price'].values

encoder = OneHotEncoder()
X = encoder.fit_transform(X)
```

More than two categories - School of Painting

The encoded data puts a one in a location representing each category in the data

```
['A', 'D/FL', 'F', 'G', 'I', 'S', 'X']
```

so

```
[1,0,0,0,0,0,0] = A
```

More than two categories - School of Painting

A [1. 0. 0. 0. 0. 0. 0.]

D/FL [0. 1. 0. 0. 0. 0. 0.]

F [0. 0. 1. 0. 0. 0. 0.]

G [0. 0. 0. 1. 0. 0. 0.]

I [0. 0. 0. 0. 1. 0. 0.]

S [0. 0. 0. 0. 0. 1. 0.]

X [0. 0. 0. 0. 0. 0. 1.]

More than two categories - School of Painting

```
LM = LinearRegression()
LM.fit(X, y)

print(LM.coef_)
print(LM.intercept_)
```

[-607.0226627 462.24135005 -265.17807213 -650.46710718
-304.23294435 2045.54876584 -680.88932952]

715.0226600517791

More than two categories - School of Painting

What the heck does this mean?!?!?

Label	Coef
A	-607.0226627
D/FL	462.24135005
F	-265.17807213
G	-650.46710718
I	-304.23294435
S	2045.54876584
X	-680.88932952

We we can interpret this as the average base cost of a painting was about 715.02, but if the painting was Austrian then on average it sold from 607.02 less but if it was Spanish then on average it sold for 2045.55 more.