Regression: Predicting House Prices

Emily Fox
University of Washington
March 27, 2018

Predicting house prices
How much is my house worth?

I want to list my house for sale

How much is my house worth?

$$ $$ $$$$
Data

\[ (x_1 = \text{sq.ft.}, \ y_1 = 
\]
\[ (x_2 = \text{sq.ft.}, \ y_2 = 
\]
\[ (x_3 = \text{sq.ft.}, \ y_3 = 
\]
\[ (x_4 = \text{sq.ft.}, \ y_4 = 
\]
\[ (x_5 = \text{sq.ft.}, \ y_5 = 
\]

\vdots

Input vs. Output:
- \( y \) is the quantity of interest
- Assume \( y \) can be predicted from \( x \)

Look at recent sales in my neighborhood

How much did they sell for?
Plot recent house sales (Past 2 years)

**Terminology:**
- \( x \) – feature, covariate, or predictor
- \( y \) – observation or response

Predict your house by similar houses

No house sold recently had *exactly* the same sq.ft.
Predict your house by similar houses

- Look at average price in range
- **Still only 2 houses!**
- Throwing out info from all other sales

Model – How we *assume* the world works

Regression model:
“Essentially, all models are wrong, but some are useful.”
George Box, 1987.

Model – How we assume the world works

Training Data → Feature extraction → ML model → ŷ

ML algorithm

Quality metric

price ($) → square feet (sq.ft.) → y → x
Linear regression
Use a simple **linear** regression model

Fit a line through the data

\[
y_i = w_0 + w_1 x_i + \varepsilon_i
\]

\[f(x) = w_0 + w_1 x\]

parameters of model

Which line?

\[
f(x) = w_0 + w_1 x
\]

different parameters \(w_0, w_1\)
“Cost” of using a given line

Residual sum of squares (RSS)

\[
\text{RSS}(w_0, w_1) = \left( \text{\$house 1} - [w_0 + w_1 \text{sq.ft.\,house 1}] \right)^2 \\
+ \left( \text{\$house 2} - [w_0 + w_1 \text{sq.ft.\,house 2}] \right)^2 \\
+ \left( \text{\$house 3} - [w_0 + w_1 \text{sq.ft.\,house 3}] \right)^2 \\
+ \ldots \text{ [include all houses]}
\]
Find “best” line

Minimize cost over all possible \( w_0, w_1 \)

\[
\text{RSS}(w_0, w_1) = \sum_{i=1}^{N} (y_i - [w_0 + w_1 x_i])^2
\]

\[
\begin{align*}
\text{RSS}(w_0=1.1, w_1=0.8) \\
\text{RSS}(w_0=0.98, w_1=0.87) \\
\text{RSS}(w_0=0.97, w_1=0.85)
\end{align*}
\]
Minimizing the cost

Minimize function over all possible $w_0, w_1$

$$\min_{w_0, w_1} \sum_{i=1}^{N} (y_i - [w_0 + w_1 x_i])^2$$

$\text{RSS}(w_0, w_1)$ is a function of 2 variables

Gradient descent

Algorithm:

while not converged

$w^{(t+1)} \leftarrow w^{(t)} - \eta \nabla \text{RSS}(w^{(t)})$

$\hat{W}$
The fitted line: use + interpretation

Model vs. fitted line

Regression model:
\[ y_i = w_0 + w_1 x_i + \varepsilon_i \]

Estimated parameters:
\[ \hat{w}_0, \hat{w}_1 \]

\[ \hat{f}(x) = \hat{w}_0 + \hat{w}_1 x \]
Seller: Predicting your house price

\[ \hat{f}(x) = \hat{w}_0 + \hat{w}_1 x \]

Regression model:
\[ y_i = w_0 + w_1 x_i + \varepsilon_i \]

Best guess of your house price:
\[ \hat{y}_{\text{house}} = \hat{w}_0 + \hat{w}_1 \text{sq.ft.}_{\text{house}} \]

Buyer: Predicting size of house

\[ \hat{f}(x) = \hat{w}_0 + \hat{w}_1 x \]

Regression model:
\[ y_i = w_0 + w_1 x_i + \varepsilon_i \]

Best guess of size of house you can afford:
\[ \$_{\text{in bank}} = \hat{w}_0 + \hat{w}_1 \text{sq.ft.} \]
A concrete example

\[ \hat{f}(x) = -44850 + 280.76x \]

Predict $ of 2,640 sq.ft. house:
\[ -44850 + 280.76 \times 2,640 = $696,356 \]

Predict sq.ft. of $859,000 sale:
\[ \frac{(859000+44850)}{280.76} = 3,219 \text{ sq.ft.} \]

Interpreting the coefficients

\[ \hat{y} = \hat{w}_0 + \hat{w}_1 x \]

Predicted $ of house with sq.ft. = 0 (just land)
Interpreting the coefficients

\[ \hat{y} = \hat{w}_0 + \hat{w}_1 x \]

**Warning:** magnitude depends on units of both features and observations
A concrete example

\[
\hat{f}(x) = -44,850 + 280.76 \text{ ($/sq.ft.$)} x
\]

Predict $ of 2,640 sq.ft. house:

\[
-44,850 + 280.76 \times 2,640 = 696,356
\]

Predict sq.ft. of $859,000 sale:

\[
\frac{859,000 + 44,850}{280.76} = 3,219 \text{ sq.ft.}
\]

But what if:
- House was measured in square meters?
- Price was measured in RMB?
Adding higher order effects

Fit data with a line or ... ?

You show your friend your analysis
Fit data with a line or ... ?

Dude, it’s not a linear relationship!

What about a quadratic function?

Dude, it’s not a linear relationship!
What about a quadratic function?

\[ f_w(x) = w_0 + w_1 x + w_2 x^2 \]

Even higher order polynomial

\[ f(x) = w_0 + w_1 x + w_2 x^2 + \ldots + w_p x^p \]
Polynomial regression

Model:
\[ y_i = w_0 + w_1 x_i + w_2 x_i^2 + \ldots + w_p x_i^p + \varepsilon_i \]

- treat as different features
- feature 1 = 1 (constant)  \quad parameter 1 = w_0
- feature 2 = x  \quad parameter 2 = w_1
- feature 3 = x^2  \quad parameter 3 = w_2
- \ldots
- feature p+1 = x^p  \quad parameter p+1 = w_p

Generic basis expansion

Model:
\[ y_i = w_0 h_0(x_i) + w_1 h_1(x_i) + \ldots + w_D h_D(x_i) + \varepsilon_i \]
\[ = \sum_{j=0}^{D} w_j h_j(x_i) + \varepsilon_i \]
- j^{th} feature
- j^{th} regression coefficient or weight
Generic basis expansion

Model:
\[ y_i = w_0 h_0(x_i) + w_1 h_1(x_i) + \ldots + w_D h_D(x_i) + \epsilon_i \]
\[ = \sum_{j=0}^{D} w_j h_j(x_i) + \epsilon_i \]

feature 1 = \( h_0(x) \)…often 1 (constant)
feature 2 = \( h_1(x) \)… e.g., \( x \)
feature 3 = \( h_2(x) \)… e.g., \( x^2 \) or \( \sin(2\pi x/12) \) or \( \log(x) \)
...
feature \( D+1 \) = \( h_D(x) \)… e.g., \( x^p \)

Adding other features
Predictions just based on house size

\[
f(x) = w_0 + w_1 \text{sq.ft.} + w_2 \#\text{bath}
\]

Add more inputs

Only 1 bathroom! Not same as my 3 bathrooms
Many possible inputs

- Square feet
- # bathrooms
- # bedrooms
- Lot size
- Year built
- ...

General notation

Output: $y$  

Inputs: $x = (x[1], x[2], \ldots, x[d])$  

Notational conventions:
- $x[j]$ = $j^{th}$ input ($scalar$)  
- $h_j(x)$ = $j^{th}$ feature ($scalar$)  
- $x_i$ = input of $i^{th}$ data point ($vector$)  
- $x_i[j]$ = $j^{th}$ input of $i^{th}$ data point ($scalar$)
Generic linear regression model

Model:
\[ y_i = w_0 h_0(x_i) + w_1 h_1(x_i) + \ldots + w_D h_D(x_i) + \epsilon_i = \sum_{j=0}^{D} w_j h_j(x_i) + \epsilon_i \]

- feature 1 = \( h_0(x) \) ... e.g., 1
- feature 2 = \( h_1(x) \) ... e.g., \( x[1] \) = sq. ft.
- feature 3 = \( h_2(x) \) ... e.g., \( x[2] \) = #bath
  - or, log(\( x[7] \)) \( x[2] \) = log(#bed) x #bath
- ...
- feature \( D+1 = h_D(x) \) ... some other function of \( x[1], \ldots, x[d] \)

More on notation

- # observations \( (x_i, y_i) : N \)
- # inputs \( x[j] : d \)
- # features \( h_j(x) : D \)
RSS for multiple regression

\[ RSS(w) = \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 \]

How many features to use?

- More on this soon!
A compact representation

Representing function using vectors

\[ f(x) = w_0 h_0(x) + w_1 h_1(x) + \ldots + w_D h_D(x) = \sum_{j=0}^{D} w_j h_j(x) \]
As an algorithm...

\[
f(x) = w_0 h_0(x) + w_1 h_1(x) + \ldots + w_D h_D(x) = \sum_{j=0}^{D} w_j h_j(x)
\]

Algorithm:

\[
\text{for } j=0,\ldots,D
\]

Compact notation

\[
f(x) = w_0 h_0(x) + w_1 h_1(x) + \ldots + w_D h_D(x) = \sum_{j=0}^{D} w_j h_j(x)
\]

\[
\begin{align*}
1 & 0 & 0 & 0 & 5 & 3 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
3 & 0 & 0 & 0 & 2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
\end{align*}
\]
Interpreting the fitted function

Interpreting the coefficients – Simple linear regression

\[ \hat{y} = \hat{w}_0 + \hat{w}_1 x \]

- Predicted change in 
- 1 sq. ft.

price (\$)  

square feet (sq.ft.)
Interpreting the coefficients – Two linear features

\[ \hat{y} = \hat{w}_0 + \hat{w}_1 x[1] + \hat{w}_2 x[2] \]

For fixed \# sq.ft.!

For fixed # bathrooms

predicted change in $
Interpreting the coefficients –

**Multiple linear features**

\[
\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \ldots + \hat{w}_j x_j + \ldots + \hat{w}_d x_d
\]

- Square feet (sq.ft.)
- Price ($)
- # Bathrooms

Can’t hold other features fixed!

**Polynomial regression**

\[
\hat{y} = \hat{w}_0 + \hat{w}_1 x + \ldots + \hat{w}_j x^j + \ldots + \hat{w}_p x^p
\]

- Square feet (sq.ft.)
- Price ($)

Can’t hold other features fixed!
Summary for regression

Training Data → Feature extraction → ML model → Quality metric → ML algorithm → ŷ

Symbol explanation:
- ŷ: Predicted value
- ŷ: True value
- w: Model parameters
What you can do now...

- Describe the input (features) and output (real-valued predictions) of a regression model
- Calculate a goodness-of-fit metric (e.g., RSS)
- Understand how gradient descent is used to estimate model parameters by minimizing RSS
- Exploit the estimated model to form predictions
- Describe a regression model using multiple features
- Interpret coefficients in a regression model with multiple features
- Describe other applications where regression is useful