Linear Regression

Reading

- Chapters 3.1, 3.2
  - Hastie, Trevor, et al. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. New York: springer, 2009.
  - Available online: https://hastie.su.domains/Papers/ESLII.pdf
- Chapters 3.1.1, 3.1.2, 3.2.1, 3.2.2
  - James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: springer, 2013.
  - Available online: https://www.statlearning.com/
- Linear regression from a statistical point of view

Overview

Rensselaer

- Linear regression is one of the simplest and best understood methods in statistics/ML
- We can derive closed-form optimal solutions in many cases
- It works well with some of the fundamental results of probability theory, e.g., the Central Limit Theorem
- It has good generalization capacity for many learning problems in practice
- Most successful modern ML methods (deep learning, SVMs) are direct extensions of linear methods
- Understanding linear methods is a necessary condition for understanding more advanced topics

Linear Regression Setup

- As usual, we are given N labeled IID examples: $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$
  - where $\boldsymbol{x}_i \in \mathbb{R}^p, y_i \in \mathbb{R}$
- We assume the examples are sampled from $\mathcal{D}$ and are realizations of random variables $(\boldsymbol{X}, Y) \sim \mathcal{D}$
- The goal of linear methods is to find a linear f such that

$$Y = f(\boldsymbol{X})$$

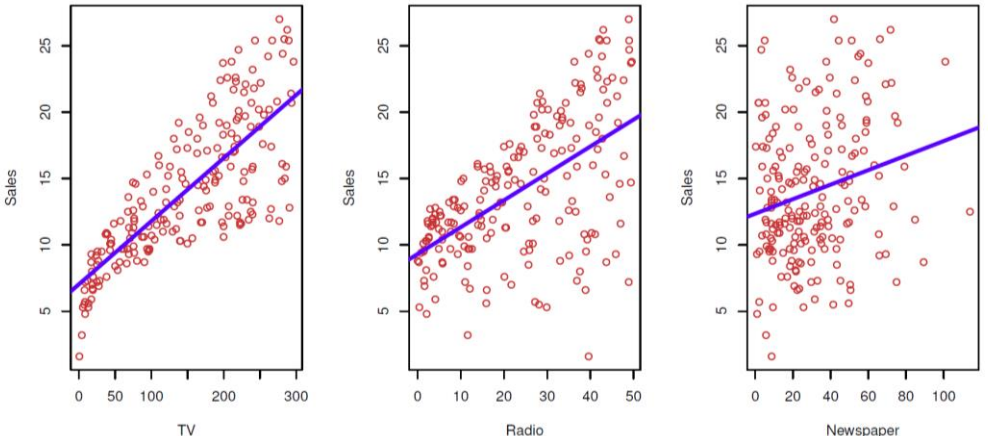
- Specifically, let $\boldsymbol{X} = \left[X_1, \ldots, X_p\right]^T$
- The goal is to find parameters w_i such that

$$Y = w_0 + w_1 X_1 + \cdots + w_p X_p$$

  - The book uses β_i for parameters but w_i is the standard notation in ML

Advertising Example

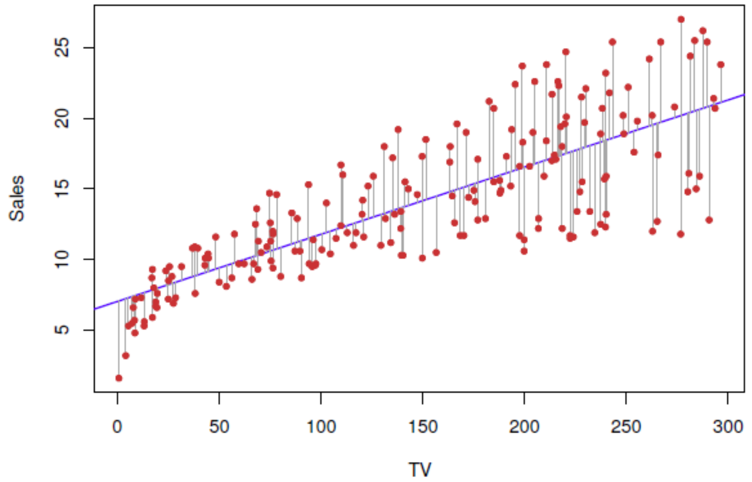
- Suppose you are a sales analyst and would like to asses the benefit of different ways of advertising
  - You have data for a product sold at 200 different markets
  - For each market, the product has been advertised on TV, radio and newspaper



- Your job is to analyze the relative contribution of each advertising method
  - You need to build a model of how ad spending affects sales

Advertising Example, cont'd

- Suppose you hypothesize that there is a linear relationship between the amount of dollars invested and the sales
  - Of course, true relationship is unknown
  - But if a line captures most of the variability (modulo some noise), then that's a good start
- Consider first just TV ads
  - How do you estimate the slope (and intercept) of the line?



Best fit line

- Suppose we want a line that minimizes the average distance to all points
  - How do we pick w_0 and w_1?
- First note that for any w_0, w_1:
  - Given an example x_i, the line prediction is $\hat{y}_i = w_0 + w_1 x_i$
  - The prediction error is

$$e_i = y_i - \hat{y}_i = y_i - (w_0 + w_1 x_1)$$

- Suppose we pick the weights to minimize $\frac{1}{N}\sum_{i=1}^{N} e_i$
  - What is wrong with this strategy?
    - e_i can be made arbitrarily negative, i.e., minimum is $-\infty$
  - What's an alternative formulation?
  - Least squares!

Least Squares

- Instead of minimizing the average error, minimize the sum of squared errors

$$\frac{1}{N}\sum_{i=1}^{N} e_i^2 =$$

$$= \frac{1}{N}\sum_{i=1}^{N}(y_i - w_0 - w_1 x_i)^2$$

  – The problem is now well defined because $e_i^2 \geq 0$
  – This approach has many names: least squares, minimum squared error (MSE), residual sum of squares
  – Note that $1/N$ is constant and doesn't affect the w_0 and w_1 that minimize the MSE

Minimize the sum of squares

- First, consider the special case $w_0 = 0$
- Problem is

$$\min_{w_1} \sum_{i=1}^{N} (y_i - w_1 x_i)^2$$

- Expanding the parentheses, we get

$$\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N} 2 w_1 y_i x_i + \sum_{i=1}^{N} w_1^2 x_i^2$$

- Quadratic equation in w_1, min is achieved when derivative is 0

Minimize the sum of squares, cont'd

- Expanding the parentheses, we get

$$\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N} 2w_1 y_i x_i + \sum_{i=1}^{N} w_1^2 x_i^2$$

- Quadratic equation in w_1, min is achieved when derivative is 0
- Derivative w.r.t w_1 is

$$-2\sum_{i=1}^{N} y_i x_i + 2w_1 \sum_{i=1}^{N} x_i^2 = 0$$

$$w_1^* = \frac{\sum_{i=1}^{N} y_i x_i}{\sum_{i=1}^{N} x_i^2}$$

- If we stack all data in vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, then $w_1^* = \frac{\boldsymbol{y}^T \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}$

What about multiple dimensions?

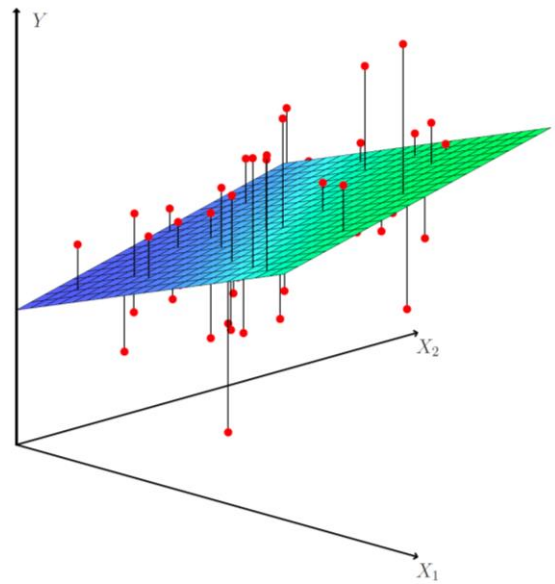
- What if you would like to build a model that takes all 3 X variables as inputs, i.e.,

$$f(\boldsymbol{X}) = w_0 + w_1 X_1 + w_2 X_2 + w_3 X_3$$

- Why would you do this instead of building a separate model for each dimension?
  - Can capture interactions between different dimensions
  - E.g., suppose TV ads are the most effective, but all ads were increased simultaneously
    - In each dimension, there will be a *correlation* between ad spending and sales
    - But if you build the 3D model, the TV coefficient will likely dominate
    - In general, causality is very hard to capture, but building a multi-dimensional model is always better than building many 1D models

Multiple dimensions, cont'd

- The function f now becomes a plane
- Individual regression coefficients



	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

- Multiple-dimension regression coefficients
  - Note the newspaper coefficient

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Interpretation of linear coefficients

- Suppose we have obtained a parameter estimate $\widehat{w}_1$
- We say that a unit change in X_1 is correlated, on average, with a $\widehat{w}_1$ unit change in Y
  - Note the word "correlated"! It is very difficult to establish causality using a purely data-driven method
  - Also note the expression "on average"! The coefficient $\widehat{w}_1$ is averaged over all training points
    - will have different prediction error for different points
- This interpretation is specific to linear models
  - But causality is hard to establish in any setting!

Multidimensional regression

- Consider the multi-dimensional linear function
$$f(\boldsymbol{X}) = w_0 + w_1 X_1 + \cdots + w_p X_p$$
- Without loss of generality, we can write
$$f(\boldsymbol{X}) = \boldsymbol{w}^T \boldsymbol{X}^*$$
  – where $\boldsymbol{w} = \left[w_0\ w_1 \dots w_p\right]^T$
  – How?
  – Rewrite $\boldsymbol{X}^* = \left[1\ \ \boldsymbol{X}^T\right]^T$
  – To avoid clutter, we will just write $\boldsymbol{X}$ instead of $\boldsymbol{X}^*$

Multidimensional Least Squares

- Goal is the same as in the 1D case
  - Find $\boldsymbol{w}$ such that the line minimizes squared errors
- For a given $\boldsymbol{w}$, the prediction is $\hat{y}_i = \boldsymbol{w}^T \boldsymbol{x}_i$
  - The prediction error is

$$e_i = y_i - \hat{y}_i$$

- And the sum of squares is

$$\sum_{i=1}^{N} e_i^2 =$$

$$= \sum_{i=1}^{N} \left(y_i - \boldsymbol{w}^T \boldsymbol{x}_i\right)^2$$

Multidimensional Least Squares, cont'd

- The sum of squares is

$$\sum_{i=1}^{N} \left(y_i - \boldsymbol{w}^T \boldsymbol{x}_i\right)^2$$

- To minimize, once again expand the parentheses

$$\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N} 2 y_i \boldsymbol{w}^T \boldsymbol{x}_i + \sum_{i=1}^{N} \left(\boldsymbol{w}^T \boldsymbol{x}_i\right)^2$$

- Then, take gradient w.r.t. $\boldsymbol{w}$ and set equal to 0

$$-2 \sum_{i=1}^{N} y_i \boldsymbol{x}_i + 2 \sum_{i=1}^{N} \left(\boldsymbol{w}^T \boldsymbol{x}_i\right) \boldsymbol{x}_i = 0$$

Aside: Vector Calculus

- Suppose we are given a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$
- What is the derivative of f?
- When $n = 1$, it is just the partial derivative $f' = \frac{\partial f}{\partial x}$
- When $n > 1$, the derivative is a vector of all partial derivatives:

$$\nabla_{\boldsymbol{x}} f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \cdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

  – This is called the gradient of f
  – The gradient is the multi-dimensional extension of the derivative

Multidimensional Least Squares, cont'd

$$-2\sum_{i=1}^{N} y_i \boldsymbol{x}_i + 2\sum_{i=1}^{N} (\boldsymbol{w}^T \boldsymbol{x}_i)\boldsymbol{x}_i = 0$$

- Temporary notation: Let $\boldsymbol{y} = [y_1, \ldots y_N]^T$ and $\boldsymbol{X} = [\boldsymbol{x}_1 \ldots \boldsymbol{x}_N]$
- Note that for any matrix $\boldsymbol{A}$ and vector $\boldsymbol{x}$, the following is true
$$\boldsymbol{A}\boldsymbol{x} = x_1 \boldsymbol{a}_1 + \cdots + x_n \boldsymbol{a}_n$$
- Thus, $\sum_{i=1}^{N} y_i \boldsymbol{x}_i = \boldsymbol{X}\boldsymbol{y}$
- Similarly, $\sum_{i=1}^{N} (\boldsymbol{w}^T \boldsymbol{x}_i)\boldsymbol{x}_i = \boldsymbol{X}\begin{bmatrix} \boldsymbol{w}^T \boldsymbol{x}_1 \\ \ldots \\ \boldsymbol{w}^T \boldsymbol{x}_N \end{bmatrix} = \boldsymbol{X}(\boldsymbol{w}^T \boldsymbol{X})^T = \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{w}$

Multidimensional Least Squares, cont'd

$$-2\sum_{i=1}^{N} y_i \boldsymbol{x}_i + 2\sum_{i=1}^{N} \left(\boldsymbol{w}^T \boldsymbol{x}_i\right)\boldsymbol{x}_i = 0$$

- Temporary notation: Let $\boldsymbol{y} = [y_1, \dots y_N]^T$ and $\boldsymbol{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_N]$
- Then the above becomes

$$-2\boldsymbol{X}\boldsymbol{y} + 2\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{w} = 0$$

$$\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{w} = \boldsymbol{X}\boldsymbol{y}$$

- To solve for $\boldsymbol{w}$, we need to multiply by $\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}$ on the left
  - When is that matrix invertible?
  - Recall $\boldsymbol{X} \in \mathbb{R}^{p+1 \times N}$, so $\boldsymbol{X}$ must be a wide matrix
  - Typically, we need much more examples than dimensions for learning to succeed (i.e., $N \gg p$)

Multidimensional Least Squares, cont'd

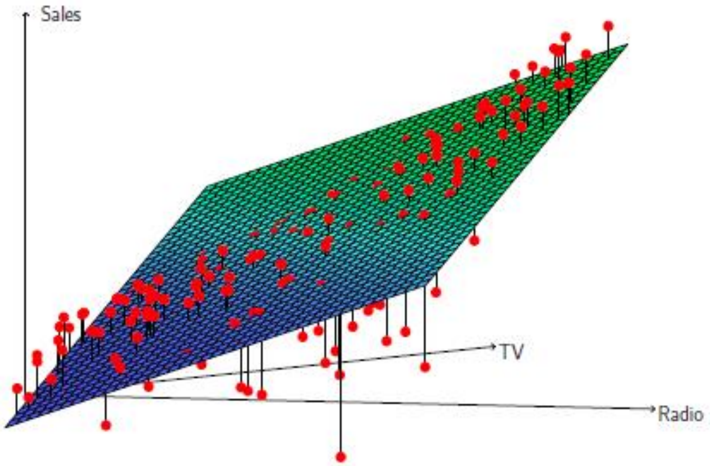
$$\boldsymbol{w}^* = \left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{y}$$

- Notice that $\boldsymbol{X}\boldsymbol{X}^T$ is symmetric
  - Why?

$$\left(\boldsymbol{X}\boldsymbol{X}^T\right)^T = \left(\boldsymbol{X}^T\right)^T\boldsymbol{X}^T = \boldsymbol{X}\boldsymbol{X}^T$$

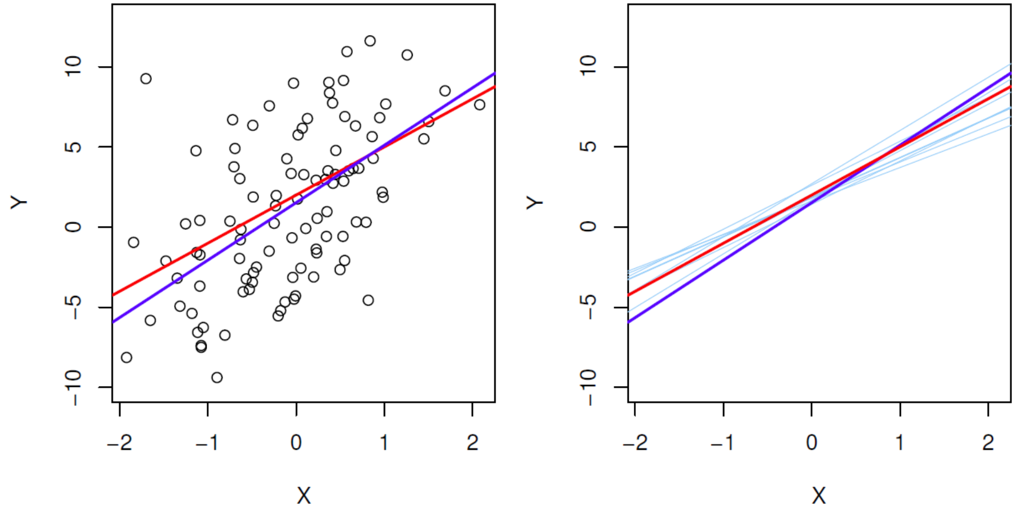
How accurate are our parameter estimates?

- There are several factors to consider when talking about accuracy
- Is the true relationship linear or close to linear?
  - If not, then no line will be a great predictor
- In many real-life cases, relationship is not truly linear but a linear model is still a good way to describe trends



How accurate are our parameter estimates?

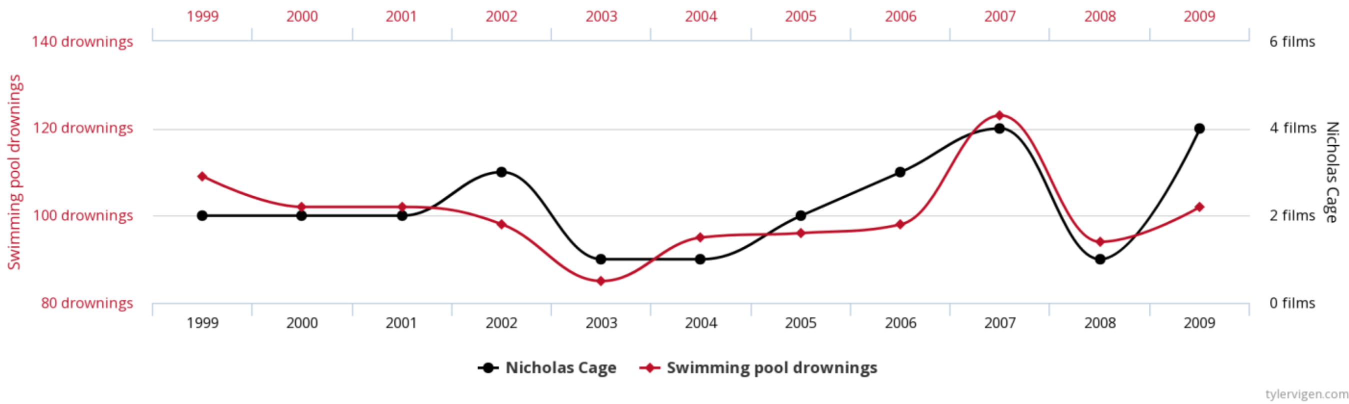
- If the relationship is linear, how close to the true line is the line we learned?



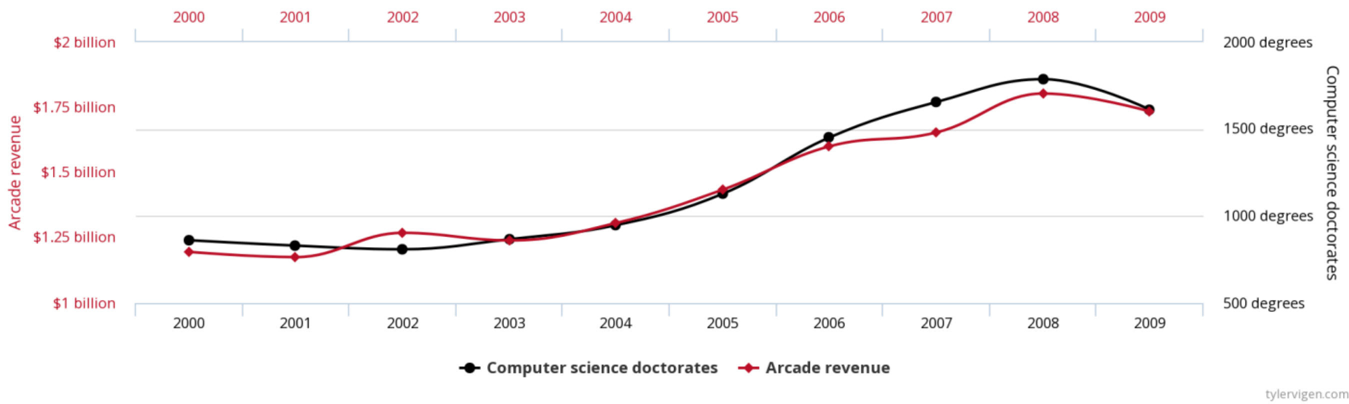
  - Red: true line; Blue: learned lines
  - As we collect more data, the learned line will converge to the true line (Law of Large Numbers)
  - Each slope estimate follows a bell-shaped distribution
    - Converges to a Gaussian with more data (Central Limit Theorem)

Spurious Correlation Examples

Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in



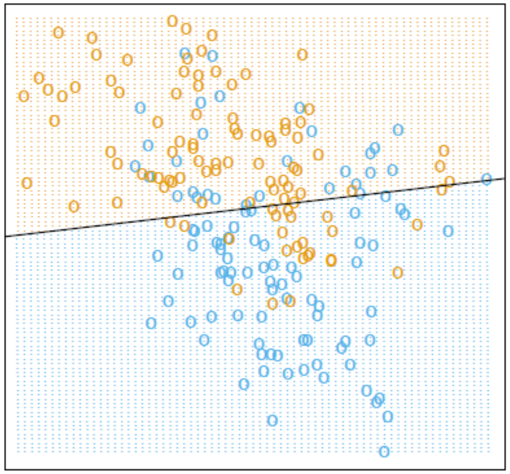
Total revenue generated by arcades
correlates with
Computer science doctorates awarded in the US



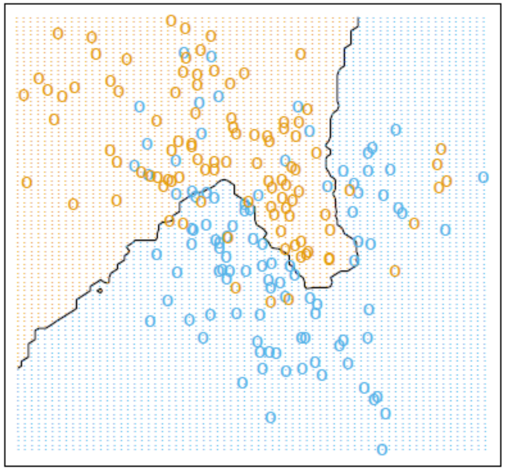
source: http://www.tylervigen.com/spurious-correlations

Linear Regression vs Nearest Neighbors

- Linear regression cannot capture complex relationships
  - Will talk more about classification next
- Nearest neighbor actually works quite well in some cases
  - What are cases where nearest neighbor would not work so well?
  - High-dimensional settings where data is sparse
  - This issue is called the curse of dimensionality
    - Quite common across ML



15-Nearest Neighbor Classifier



First Dataset: MNIST

- A dataset of 60K grayscale images of handwritten digits

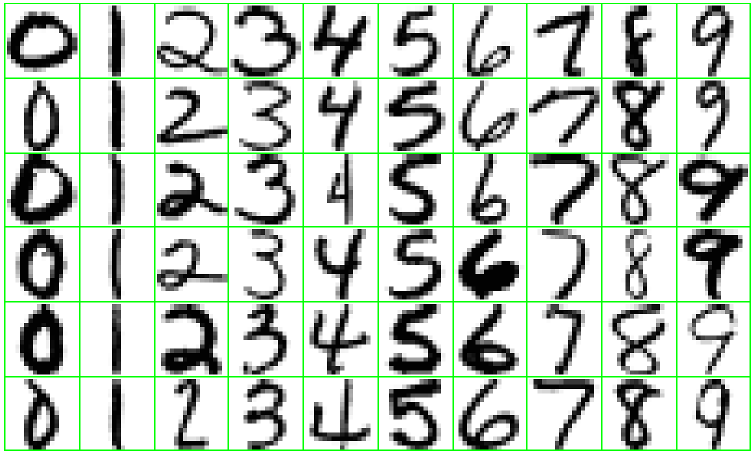


FIGURE 1.2. *Examples of handwritten digits from U.S. postal envelopes.*

- Each image is a 28×28 matrix of pixels
  - Each pixel is an integer between 0 and 255
  - Often normalized between 0 and 1 for numeric stability
- Dataset more or less *solved*
  - Can achieve >99% accuracy with various methods

MNIST, cont’d

- We’ll have a few homeworks on MNIST
- First, we’ll try a linear regression/classifiication method
  - Does this make sense?
  - What do you expect to see?

What about non-linear terms?

- One can add non-linear terms to the function f, e.g.,
$$f(\boldsymbol{X}) = w_0 + w_1 X_1 + w_2 X_2 + w_3 X_1 X_2$$
- And then learn the coefficients in the same way using MSE
- You should only do this if you have a good reason to believe this non-linearity is present in the data
- Intuitive interpretation gets harder for non-linear models
  - In general, non-linear models complicate the math very quickly, and statistical guarantees are harder to get
- In modern ML, if the data has an unknown non-linear relationship, then neural networks are the model of choice
  - More on this later