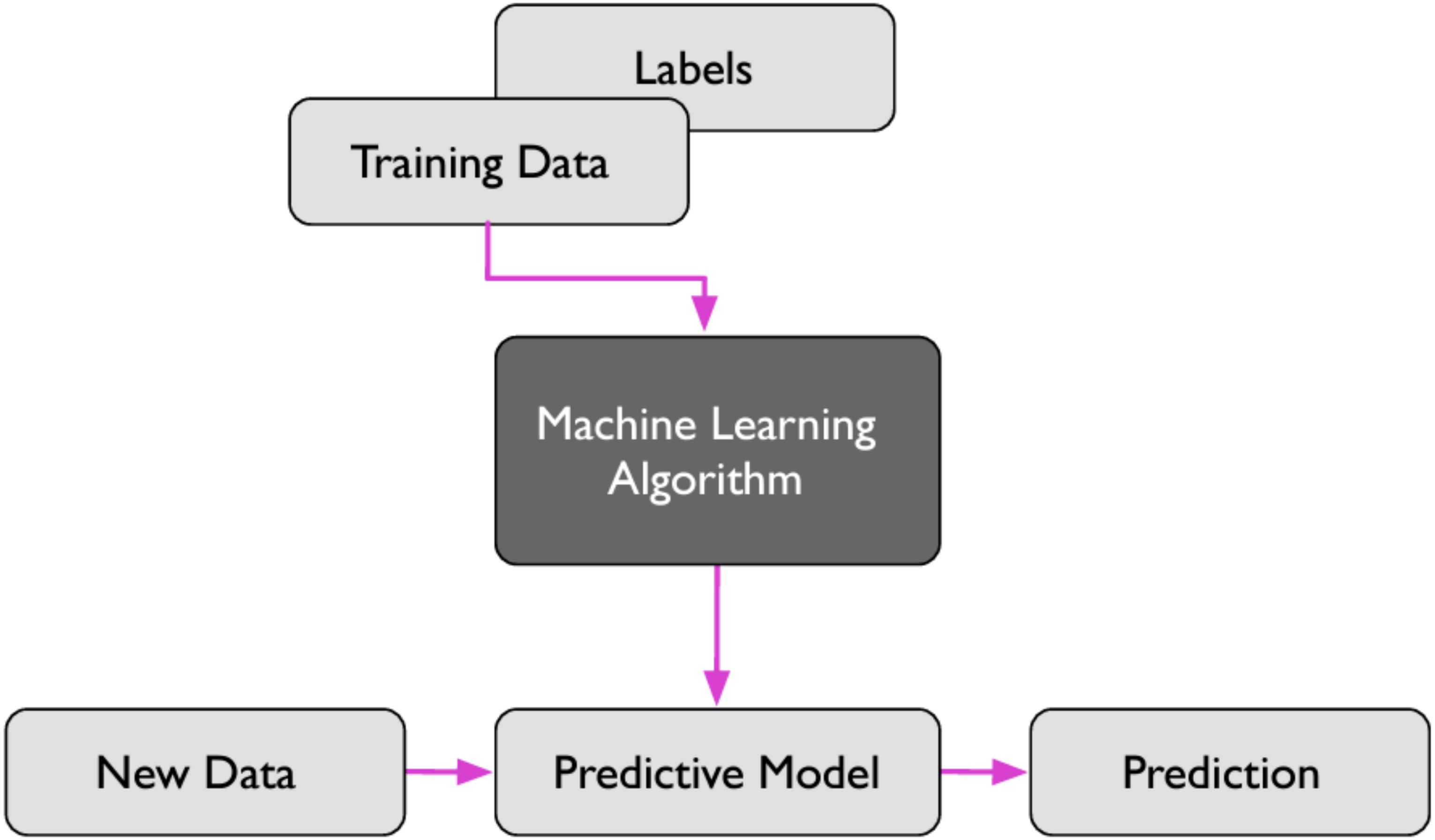
LECTURE 3: MULTIPLE VARIABLE LINEAR REGRESSION

Ehsan Aryafar
earyafar@pdx.edu
http://web.cecs.pdx.edu/~aryafare/ML.html

Recall: Supervised Learning

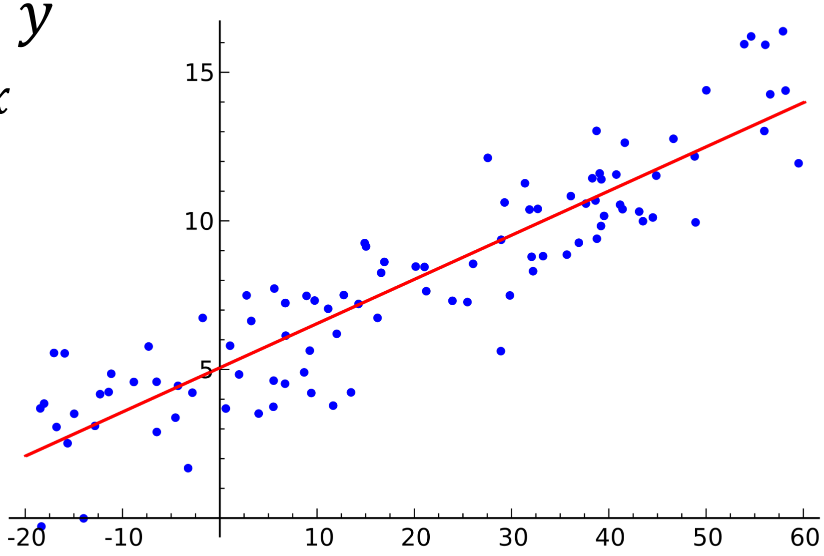


Recall: Linear Model Residual

- Knowing x does not exactly predict y
  - Variation in y due to factors other than x
- Add a residual term

$$y = \beta_0 + \beta_1 x + \epsilon$$

- Residual = component the model does not explain
  - Predicted value: $\hat{y}_i = \beta_1 x_i + \beta_0$
  - Residual: $\epsilon_i = y_i - \hat{y}_i$
  - *Sort of like residual error*
- Vertical deviation from the regression line



Recall: Least Squares Model Fitting

- How do we select parameters $\beta = (\beta_0, \beta_1)$?
- Define $\hat{y}_i = \beta_1 x_i + \beta_0$
  - Predicted value on sample i for parameters $\beta = (\beta_0, \beta_1)$
- Define average residual sum of squares:

$$\mathrm{RSS}(\beta_0, \beta_1) := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

  - Note that $\hat{y}_i$ is implicitly a function of $\beta = (\beta_0, \beta_1)$
  - Also called the sum of squared residuals (SSR) and sum of squared errors (SSE)
- Least squares solution: Find (β_0, β_1) to minimize RSS.
  - Geometrically, minimizes squared distances of samples to regression line

Recall: Minimizing RSS

- To minimize $\mathrm{RSS}(\beta_0, \beta_1)$ take partial derivatives:

$$\frac{\partial \mathrm{RSS}}{\partial \beta_0} = 0, \qquad \frac{\partial \mathrm{RSS}}{\partial \beta_1} = 0$$

- Taking derivatives we get two conditions (proof removed):

$$\sum_{i=1}^{N} \epsilon_i = 0, \qquad \sum_{i=1}^{N} x_i \epsilon_i = 0 \quad \text{where} \quad \epsilon_i = y_i - \beta_0 - \beta_1 x_i$$

- Regression equation:
  - After some manipulation, solution to optimal slope and intercept:

$$\beta_1 = \frac{s_{xy}}{s_x^2} = \frac{r_{xy} s_y}{s_x}, \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

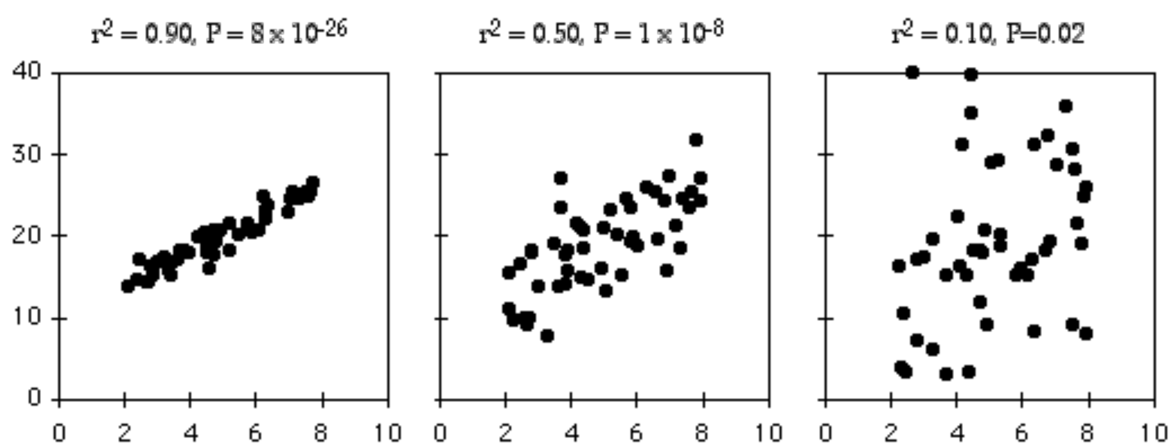
Correlation coefficient between x and y

Recall: Minimum RSS

- Minimum RSS (Proof removed)

$$\min_{\beta_0,\beta_1} \mathrm{RSS}(\beta_0,\beta_1) = N\left(1 - r_{xy}^2\right)s_y^2$$

- Coefficient of Determination: $R^2 = r_{xy}^2$
  - R^2 close to one makes RSS Close to 0
  - Explains portion of variance in y explained by x
  - s_y^2=variance in target y
  - $(1 - R^2)s_y^2$=residual sum of squares after accounting for x



Learning Objectives

- Formulate a machine learning model as a multiple variable linear regression model.
  - Identify prediction vector and target for the problem.
- Write the regression model in matrix form. Write the feature matrix
- Compute the least-squares solution for the regression coefficients on training data.
- Derive the least-squares formula from minimization of the RSS

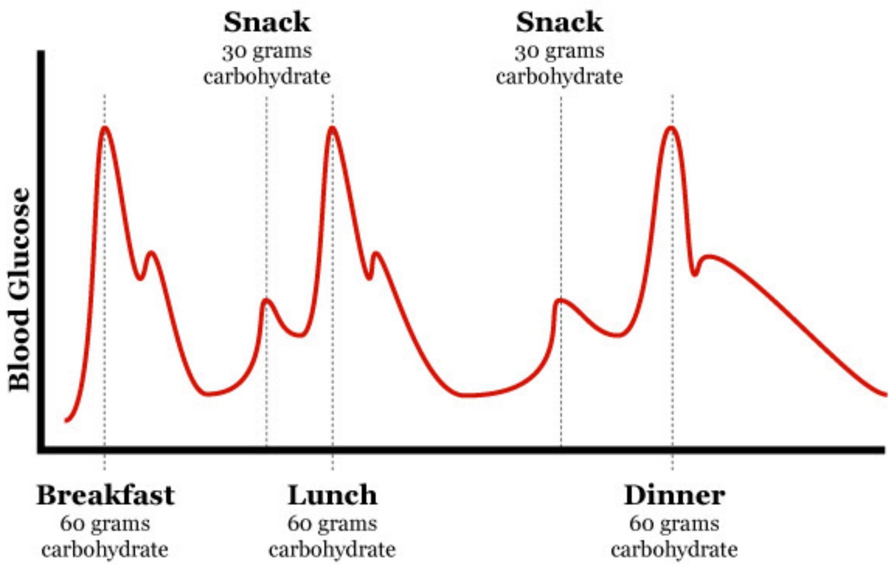
- Manipulate 2D arrays in python (indexing, stacking, computing shapes, …)
- Compute the LS solution using python linear algebra and machine learning packages

Outline

- Motivating Example: Understanding glucose levels in diabetes patients
- Multiple variable linear models
- Least squares solutions
- Computing the solutions in python
- Special case: Simple linear regression
- Extensions

Example: Blood Glucose Level

- Diabetes patients must monitor glucose level
- What causes blood glucose levels to rise and fall?
- Many factors
- We know mechanisms qualitatively
- But, quantitative models are difficult to obtain
  - Hard to derive from first principles
  - Difficult to model physiological process precisely
- Can machine learning help?



Data from AIM 94 Experiment

Diabetes Data Set

Download: Data Folder, Data Set Description

Abstract: This diabetes dataset is from AIM '94

Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	N/A	**Area:**	Life
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	20	**Date Donated**	N/A
Associated Tasks:	N/A	**Missing Values?**	N/A	**Number of Web Hits:**	161379

- Data collected as series of events
  - Eating
  - Exercise
  - Insulin dosage
- Target variable (glucose level) monitored

Data Set Information:

Diabetes patient records were obtained from two sources: ar
clock to timestamp events, whereas the paper records only p
assigned to breakfast (08:00), lunch (12:00), dinner (18:00), a
records have more realistic time stamps.

Diabetes files consist of four fields per record. Each field is s

File Names and format:
(1) Date in MM-DD-YYYY format
(2) Time in XX:YY format
(3) Code
(4) Value

The Code field is deciphered as follows:

33 = Regular insulin dose
34 = NPH insulin dose
35 = UltraLente insulin dose
48 = Unspecified blood glucose measurement
57 = Unspecified blood glucose measurement
58 = Pre-breakfast blood glucose measurement
59 = Post-breakfast blood glucose measurement
60 = Pre-lunch blood glucose measurement
61 = Post-lunch blood glucose measurement
62 = Pre-supper blood glucose measurement
63 = Post-supper blood glucose measurement

Demo

Demo: Predicting Glucose Levels using Mulitple Linear Regression

In this demo, you will learn how to:

- Fit multiple linear regression models using python's `sklearn` pachage.
- Split data into training and test.
- Manipulate and visualize multivariable arrays.

We first load the packages as usual.

[5]:
```
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline
```

Loading the Data

```python
from sklearn import datasets, linear_model, preprocessing

# Load the diabetes dataset
diabetes = datasets.load_diabetes()
X = diabetes.data
y = diabetes.target
```

```python
print(diabetes.DESCR)
```

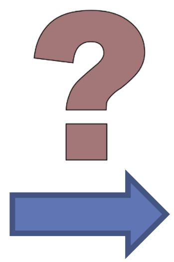
```
.. _diabetes_dataset:

Diabetes dataset
----------------

Ten baseline variables, age, sex, body mass index, average blood
pressure, and six blood serum measurements were obtained for each of n =
442 diabetes patients, as well as the response of interest, a
quantitative measure of disease progression one year after baseline.

**Data Set Characteristics:**

  :Number of Instances: 442

  :Number of Attributes: First 10 columns are numeric predictive values

  :Target: Column 11 is a quantitative measure of disease progression one year after baseline

  :Attribute Information:
      - Age
      - Sex
      - Body mass index
      - Average blood pressure
      - S1
      - S2
      - S3
      - S4
      - S5
      - S6
```

- Sklearn package:
  - Many methods for ML
  - Datasets
  - Will use throughout this class
- Diabetes dataset is one example

Finding a Mathematical Model

Attributes

x_1: Age
x_2: Sex
x_3: BMI
x_4: BP
x_5: S1
⋮
x_{10}: S6



Target
y =Glucose level

$$y \approx \hat{y} = f(x_1, \dots, x_{10})$$

- Goal: Find a function to predict glucose level from the 10 attributes
- Problem: Several attributes
  - Need a multi-variable function

Matrix Representation of Data

- Data is a matrix and a target vector
- n samples: One sample per row
- k features / attributes /predictors:
  - One feature per column
- This example:
  - y_i = blood glucose measurement of i-th sample
  - $x_{i,j}$: j-th feature of i-th sample
  - $\boldsymbol{x}_i^T = [x_{i,1}, x_{i,2}, \ldots, x_{i,k}]$: **feature or predictor vector**
  - i-th sample contains $\boldsymbol{x}_i, y_i$

Attributes

Target vector

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Samples

Outline

- Motivating Example: Understanding glucose levels in diabetes patients
- Math background
- Multiple variable linear models
- Least squares solutions
- Computing the solutions in python
- Special case: Simple linear regression
- Extensions

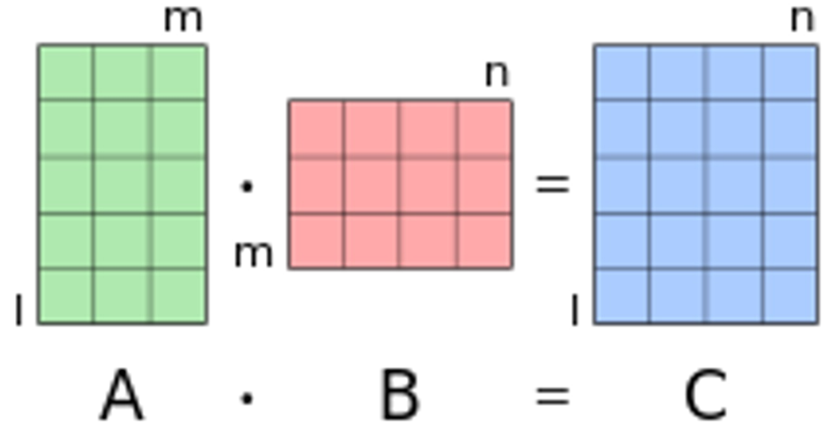
Matrix Addition

$$\mathbf{A}+\mathbf{B} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2n}+b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \cdots & a_{mn}+b_{mn} \end{bmatrix}$$

$$\begin{bmatrix} 1 & 3 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 7 & 5 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1+0 & 3+0 \\ 1+7 & 0+5 \\ 1+2 & 2+1 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 8 & 5 \\ 3 & 3 \end{bmatrix}$$

Matrix Multiplication



Matrix Multiplication

$$c_{11} = a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} + a_{14}b_{41}$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \\ b_{41} & b_{42} & b_{43} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{bmatrix}$$

2 x 4 4 x 3 2 x 3

$$c_{22} = a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} + a_{24}b_{42}$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \\ b_{41} & b_{42} & b_{43} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{bmatrix}$$

Example Matrix Multiplication

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \text{x} \begin{bmatrix} 10 & 11 \\ 20 & 21 \\ 30 & 31 \end{bmatrix}$$

$$= \begin{bmatrix} 1\text{x}10 + 2\text{x}20 + 3\text{x}30 & 1\text{x}11 + 2\text{x}21 + 3\text{x}31 \\ 4\text{x}10 + 5\text{x}20 + 6\text{x}30 & 4\text{x}11 + 5\text{x}21 + 6\text{x}31 \end{bmatrix}$$

$$= \begin{bmatrix} 10+40+90 & 11+42+93 \\ 40+100+180 & 44+105+186 \end{bmatrix} = \begin{bmatrix} 140 & 146 \\ 320 & 335 \end{bmatrix}$$

Matrix Transpose

- Transpose of a matrix A is shown as A^T
- Formally, the i-th row, j-th column element of A^T is the j-th row, i-th column element of A. Below are some examples:

$$\begin{bmatrix} 1 & 2 \end{bmatrix}^T = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}^T = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$$

Dot Product

- Algebraic definition

The dot product of two vectors **a** = [a_1, a_2, …, a_n] and **b** = [b_1, b_2, …, b_n] is defined as:

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

$$[1, 3, -5] \cdot [4, -2, -1] = (1 \times 4) + (3 \times -2) + (-5 \times -1) = 3$$

- Geometric definition (the same value as algebraic)

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos\theta$$

$$\|\mathbf{a}\| = \sqrt{\mathbf{a} \cdot \mathbf{a}}$$

Matrix Inversion

- In linear algebra, an n-by-n square matrix is called invertible, if there exists an n-by-n square matrix B such that

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$$

$$I_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

I_n is the identity matrix

- Inverse of **A** is typically denoted by $\mathbf{A}^{-1}$

Outline

- Motivating Example: Understanding glucose levels in diabetes patients
- Math background
- Multiple variable linear models
- Least squares solutions
- Computing the solutions in python
- Special case: Simple linear regression
- Extensions

Multivariable Linear Model for Glucose

Attributes

Age, Sex, BMI,BP,S1, …, S6

$\boldsymbol{x} = [x_1, \dots, x_{10}]$

$y \approx \hat{y} = f(x_1, \dots, x_{10})$

Target

y =Glucose level

- Goal: Find a function to predict glucose level from the 10 attributes
- Linear Model: Assume glucose is a linear function of the predictors:

$$[\text{glucose}] \approx [\text{prediction}] = \beta_0 + \beta_1[\text{Age}] + \cdots + \beta_4[BP] + \beta_5[\text{S1}] + \cdots + \beta_{10}[\text{S6}]$$

- General form:

$$\underbrace{y}_{\text{Target}} \approx \hat{y} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1 x_1 + \cdots + \beta_4 x_4 + \beta_5 x_5 + \cdots + \beta_{10} x_{10}}_{\text{10 Features}}$$

Multiple Variable Linear Model

- Vector of features: $\boldsymbol{x} = [x_1, \ldots, x_k]$
  - k features (also known as predictors, independent variable, attributes, covariates, …)
- Single target variable y
  - What we want to predict
- Linear model: Make a prediction $\hat{y}$

$$y \approx \hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- Data for training
  - Samples are $(\boldsymbol{x}_i, y_i)$, i=1,2,…,n.
  - Each sample has a vector of features: $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{ik}]$ and scalar target y_i
- Problem: Learn the best coefficients $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]$ from the training data

Another Example: Heart Rate Increase

- Linear Model: $[\text{HR increase}] \approx \beta_0 + \beta_1[\text{mins exercise}] + \beta_2[\text{exercise intensity}]$
- Data:

Subject number	HR before	HR after	Mins on treadmill	Speed (km/hr)	Days exercise / week
123	60	90	1	5.2	3
456	80	110	2	4.1	1
789	70	130	5	3.5	2
⋮	⋮	⋮	⋮	⋮	⋮

Measuring fitness of athletes

https://www.mercurynews.com/2017/10/29/4851089/

Why Use a Linear Model?

- Many natural phenomena have linear relationship
- Predictor has small variation
  - Suppose $y = f(x)$
  - If variation of x is small around some value x_0, then

$$y \approx f(x_0) + f'(x_0)(x - x_0) = \beta_0 + \beta_1 x,$$

$$\beta_0 = f(x_0) - f'(x_0)x_0, \qquad \beta_1 = f'(x_0)$$

- Simple to compute
- Easy to interpret relation
  - Coefficient β_j indicates the importance of feature j for the target.
- Advanced: Gaussian random variables:
  - If two variables are jointly Gaussian, the optimal predictor of one from the other is linear predictor

Slopes, Intercept and Inner Products

- Model with coefficients $\boldsymbol{\beta}$: $\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
- Sometimes use weight bias version:

$$\hat{y} = b + w_1 x_1 + \cdots + w_k x_k$$

  - $b = \beta_0$: Bias or intercept
  - $\boldsymbol{w} = \boldsymbol{\beta}_{1:k} = [\beta_1, \ldots, \beta_k]$: Weights or slope vector
- Can write either with inner product:

$$\hat{y} = \beta_0 + \boldsymbol{\beta}_{1:k} \cdot \boldsymbol{x} \qquad \text{or} \qquad \hat{y} = b + \boldsymbol{w} \cdot \boldsymbol{x}$$

- Inner product:
  - $\boldsymbol{w} \cdot \boldsymbol{x} = \sum_{j=1}^{k} w_j x_j$
  - Will use alternate notation: $\mathbf{w}^T \boldsymbol{x} = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$

Matrix Form of Linear Regression

- Data: $(\boldsymbol{x}_i, y_i), i = 1, \dots, n$
- Predicted value for i-th sample: $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$
- Matrix form

$$\underbrace{\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}}_{\hat{\boldsymbol{y}} \text{ a } n \text{ predicted values}} = \underbrace{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}}_{\boldsymbol{A} \text{ a } n\times p \text{ feature matrix}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{\boldsymbol{\beta} \text{ with } p = k+1 \text{ coefficient vector}}$$

- Matrix equation: $\hat{\boldsymbol{y}} = \boldsymbol{A}\,\boldsymbol{\beta}$

In-Class Exercise

- Linear Model: $[\text{HR increase}] \approx \beta_0 + \beta_1[\text{mins exercise}] + \beta_2[\text{exercise intensity}]$
- Data:

Subject number	HR before	HR after	Mins on treadmill	Speed (km/hr)	Days exercise / week
123	60	90	1	5.2	3
456	80	110	2	4.1	1
789	70	130	5	3.5	2
⋮	⋮	⋮	⋮	⋮	⋮

Measuring fitness of athletes
https://www.mercurynews.com/2017/10/29/4851089/

- What is the feature matrix A and target vector y? What are their dimensions?
- Suppose that after training we find parameters $\boldsymbol{\beta}$ = [0 15 3]. If the initial HR is 70 bpm, what is the predicted HR after 2 mins of exercise at 5 km/hr?

Outline

- Motivating Example: Understanding glucose levels in diabetes patients
- Math background
- Multiple variable linear models
- Least squares solutions
- Computing the solutions in python
- Special case: Simple linear regression
- Extensions

Least Squares Model Fitting

- How do we select parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)$?
- Define $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$
  - Predicted value on sample i for parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)$
- Define average residual sum of squares:

$$\mathrm{RSS}(\boldsymbol{\beta}) := \sum_{I=1}^{n} (y_i - \hat{y}_i)^2$$

  - Note that $\hat{y}_i$ is implicitly a function of $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)$
  - Also called the sum of squared residuals (SSR) and sum of squared errors (SSE)
- Least squares solution: Find $\boldsymbol{\beta}$ to minimize RSS.

Variants of RSS

- Often use some variant of RSS
  - Note: these are not standard
- Residual sum of squares: $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- RSS per sample or Mean Squared Error:

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

- Normalized RSS or Normalized MSE:

$$\frac{RSS/n}{s_y^2} = \frac{MSE}{s_y^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Finding Parameters via Optimization

A general ML recipe

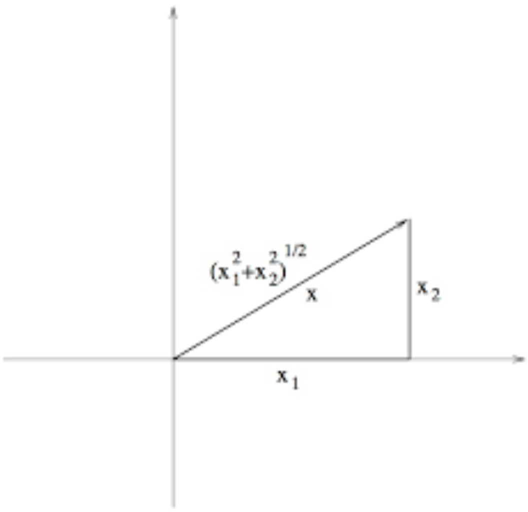
General ML problem	Multiple linear regression
Pick a model with parameters	Linear model: $\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
Get data	Data: $(\boldsymbol{x}_i, y_i), i = 1,2, \ldots, n$
Pick a loss function • Measures goodness of fit model to data • Function of the parameters	Loss function: $RSS(\beta_0, \ldots, \beta_k) \coloneqq \sum (y_i - \hat{y}_i)^2$
Find parameters that minimizes loss	Select $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)$ to minimize $RSS(\boldsymbol{\beta})$

RSS as a Vector Norm

- RSS is given by sum:

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

- Define norm of a vector:
  - $\|\boldsymbol{x}\| = (x_1^2 + \cdots + x_r^2)^{1/2}$
  - Standard Euclidean norm.
  - Sometimes called ℓ-2 norm. ℓ is for Lebesque



- Write RSS in vector form:

$$\text{RSS} = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$$

Least Squares Solution

- Consider **cost** function of the RSS:

$$\mathrm{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \qquad \hat{y}_i = \sum_{j=0}^{p} A_{ij}\beta_j$$

  - Vector $\boldsymbol{\beta}$ that minimizes RSS called the least-squares solution

- Least squares solution: The vector $\boldsymbol{\beta}$ that minimizes the RSS is:

$$\boxed{\widehat{\boldsymbol{\beta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}}$$

  - Can compute the best coefficient vector analytically
  - Just solve a linear set of equations
  - Will show the proof below

Proving the LS Formula

- Least squares formula: The vector $\boldsymbol{\beta}$ that minimizes the RSS is:

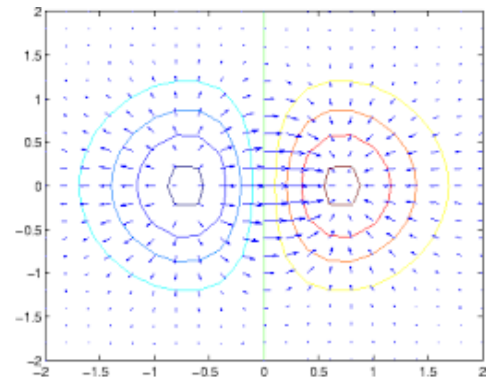
$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$$

- To prove this formula, we will:
  - Review gradients of multi-variable functions
  - Compute gradient $\nabla RSS(\boldsymbol{\beta})$
  - Solve $\nabla RSS(\boldsymbol{\beta}) = 0$

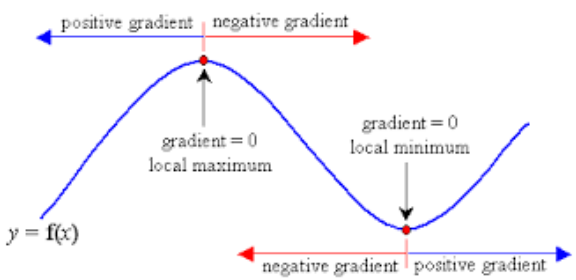
Gradients of Multi-Variable Functions

- Consider scalar valued function of a vector: $f(\boldsymbol{\beta}) = f(\beta_1, \ldots, \beta_n)$
- Gradient is the column vector:

$$\nabla f(\boldsymbol{\beta}) = \begin{bmatrix} \partial f(\boldsymbol{\beta})/\partial \beta_1 \\ \vdots \\ \partial f(\boldsymbol{\beta})/\partial \beta_n \end{bmatrix}$$



- Represents direction of maximum increase
- At a local minima or maxima: $\nabla f(\boldsymbol{\beta}) = 0$
  - Solve n equations and n unknowns
- Ex: $f(\beta_1, \beta_2) = \beta_1 \sin \beta_2 + \beta_1^2\, \beta_2$.
  - Compute $\nabla f(\boldsymbol{\beta})$.



Proof Sketch of the LS Formula

- Consider cost function of the RSS:

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \qquad \hat{y}_i = \sum_{j=0}^{p} A_{ij}\beta_j$$

  - Vector $\boldsymbol{\beta}$ that minimizes RSS called the least-squares solution
- Compute partial derivatives via chain rule: $\frac{\partial RSS}{\partial \beta_j} = -2\sum_{i=1}^{n}(y_i - \hat{y}_i)A_{ij}, = -2\sum_{i=1}^{n} A_{ij}(y_i - \hat{y}_i), \; j = 1,2,\dots,k$
- Matrix form: $\mathrm{RSS} = ||\boldsymbol{y} - A\boldsymbol{\beta}||^2, \; \nabla RSS = -2A^T(\boldsymbol{y} - \boldsymbol{A\beta})$
- Solution: $A^T(\boldsymbol{y} - A\boldsymbol{\beta}) = 0 \rightarrow \boldsymbol{\beta} = (A^T A)^{-1}A^T\boldsymbol{y}$ (least squares solution of equation $A\boldsymbol{\beta} = \mathbf{y}$)
- Minimum RSS: $RSS = \boldsymbol{y}^T[I - A(A^T A)^{-1}A^T]\boldsymbol{y}$

LS Solution via Auto-Correlation Functions

- Each data sample has a linear feature vector:
$$A_i = (A_{i0}, \cdots, A_{ik}) = (1, x_{i1}, \cdots, x_{ik})$$

- Define sample auto-correlation matrix and cross-correlation vector:
  - $R_{AA} = \frac{1}{n} A^T A,\ R_{AA}(\ell, m) = \frac{1}{n} \sum_{i=1}^{n} A_{i\ell} A_{im}$ (correlation of feature ℓ and feature *m*)
  - $R_{Ay} = \frac{1}{n} A^T y,\ R_{Ay}(\ell) = \frac{1}{n} \sum_{i=1}^{n} A_{i\ell} y_i$ (correlation of feature ℓ and target)

- Least squares solution is: $\boldsymbol{\beta} = R_{AA}^{-1} R_{Ay}$

Mean Removed Form of the LS Solution

- Often useful to remove mean from data before fitting
- Sample mean: $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, $\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$, $\bar{\boldsymbol{x}} = [\bar{x}_1, \cdots, \bar{x}_k]$
- Defined mean removed data: $\tilde{X}_{ij} = x_{ij} - \bar{x}_j, \tilde{y}_i = y_i - \bar{y}$
- Sample covariance matrix and cross-covariance vector:
  - $S_{xx}(\ell, m) = \frac{1}{N}\sum_{i=1}^{N}(x_{i\ell} - \bar{x}_\ell)(x_{im} - \bar{x}_m)$, $S_{xx} = \frac{1}{N}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$
  - $S_{xy}(\ell) = \frac{1}{N}\sum_{i=1}^{N}(x_{i\ell} - \bar{x}_\ell)(y_i - \bar{y})$, $S_{xy} = \frac{1}{N}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{y}}$
- Mean-Removed form of the least squares solution:

$$\hat{y} = \boldsymbol{\beta}_{1:k} \cdot \boldsymbol{x} + \beta_0, \qquad \boldsymbol{\beta}_{1:k} = S_{xx}^{-1} S_{xy}, \qquad \beta_0 = \bar{y} - \boldsymbol{\beta}_{1:k} \cdot \bar{\boldsymbol{x}}$$

- Proof: Removed.

R^2: Goodness of Fit

- Multiple variable coefficient of determination:

$$R^2 = \frac{s_y^2 - MSE}{s_y^2} = 1 - \frac{MSE}{s_y^2}$$

  - $\mathrm{MSE} = \frac{RSS}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - Sample variance is: $s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$,
- Interpretation:
  - $\frac{MSE}{s_y^2} = \frac{\text{Error with linear predictor}}{\text{Error predicting by mean}}$
  - R^2 = fraction of variance reduced or "explained" by the model.
- On the training data (not necessarily on the test data):
  - $R^2 \in [0,1]$ always
  - $R^2 \approx 1 \Rightarrow$ linear model provides a good fit
  - $R^2 \approx 0 \Rightarrow$ linear model provides a poor fit

Outline

- Motivating Example: Understanding glucose levels in diabetes patients
- Math background
- Multiple variable linear models
- Least squares solutions
- Computing the solutions in python
- Special case: Simple linear regression
- Extensions

Arrays and Vector in Python and MATLAB

- There are some key differences between MATLAB and Python that you need to get used to
- MATLAB
  - All arrays are at least 2 dimensions
  - Vectors are $1 \times N$ (row vectors) or $N \times 1$ (column) vectors
  - Matrix vector multiplication syntax depends if vector is on left or right: `x'*A` or `A*x`
- Python:
  - Arrays can have 1, 2, 3, … dimension
  - Vectors can be 1D arrays; matrices are generally 2D arrays
  - Vectors that are 1D arrays are neither row not column vectors
  - If `x` is 1D and `A` is 2D, then left and right multiplication are the same: `x.dot(A)` and `A.dot(x)`
- Lecture notes: We will generally treat x and x^T the same.
  - Can write $x = (x_1, \ldots, x_N)$ and still multiply by a matrix on left or right

Fitting Using sklearn

```
ns_train = 300
ns_test = nsamp - ns_train
X_tr = X[:ns_train,:]
y_tr = y[:ns_train]
```

- Return to diabetes data example
- All code provided
- Divide data into two portions:
  - Training data: First 300 samples
  - Test data: Remaining 142 samples
- Train model on training data.
- Test model (i.e. measure RSS) on test data

Calling the sklearn Linear Regression method

```
regr = linear_model.LinearRegression()
regr.fit(X_tr,y_tr)
```

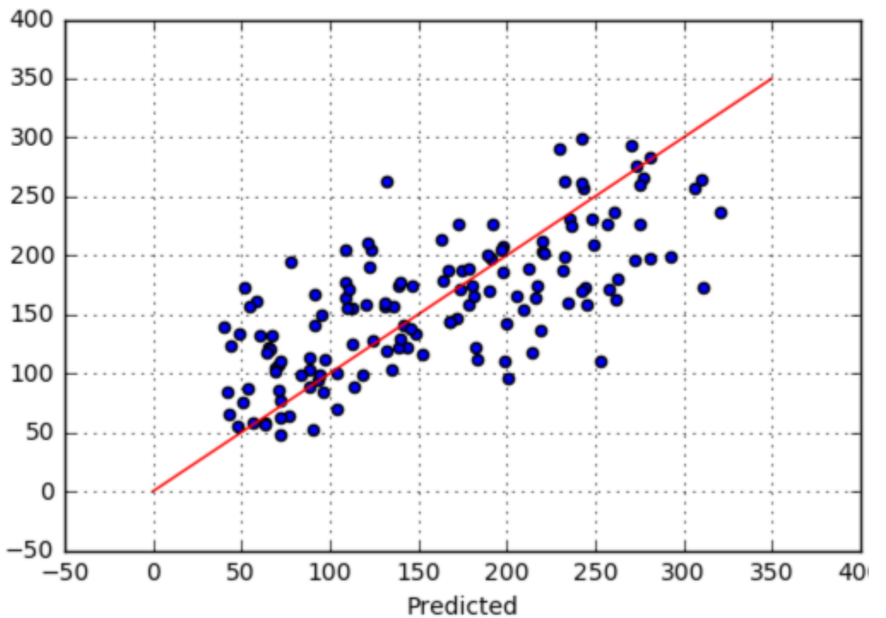
```
X_test = X[ns_train:,:]
y_test = y[ns_train:]
y_test_pred = regr.predict(X_test)
RSS_test = np.mean((y_test_pred-y_test)**2)/(np.std(y_test)**2)
Rsq_test = 1-RSS_test
print("RSS per sample = {0:f}".format(RSS_test))
print("R^2 =            {0:f}".format(Rsq_test))
```

```
RSS per sample = 0.492801
R^2 =            0.507199
```

We see that the model predicts new samples almost as well as it did the training s

```
plt.scatter(y_test,y_test_pred)
plt.plot([0,350],[0,350],'r')
plt.xlabel('Actual')
plt.xlabel('Predicted')
plt.grid()
```

- Construct a linear regression object
- Run it on the training data
- Predict values on the test data



Manually Computing the Solution

```
ones = np.ones((ns_train,1))
A = np.hstack((ones,X_tr))
```

```
out = np.linalg.lstsq(A,y_tr)
beta = out[0]
```

- Use numpy linear algebra routine to solve
  $$\beta = (A^T A)^{-1} A^T y$$
- Common mistake:
  - Compute matrix inverse $P = (A^T A)^{-1}$,
  - Then compute $\beta = P A^T y$
  - Full matrix inverse is VERY slow. Not needed.
  - Can directly solve linear system: $A\,\beta = y$
  - Numpy has routines to solve this directly

numpy.hstack function is used to stack the sequence of input arrays horizontally (i.e. column wise) to make a single array:
https://numpy.org/doc/stable/reference/generated/numpy.hstack.html

Outline

- Motivating Example: Understanding glucose levels in diabetes patients
- Math background
- Multiple variable linear models
- Least squares solutions
- Computing the solutions in python
- Special case: Simple linear regression
- Extensions

Simple vs. Multiple Regression

- Simple linear regression: One predictor (feature)
  - Scalar predictor x
  - Linear model: $\hat{y} = \beta_0 + \beta_1 x$
  - Can only account for one variable

- Multiple linear regression: Multiple predictors (features)
  - Vector predictor $\boldsymbol{x} = (x_1, \dots, x_k)$
  - Linear model: $\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
  - Can account for multiple predictors
  - Turns into simple linear regression when $k = 1$

Comparison to Single Variable Models

- We could compute models for each variable separately:

$$y = a_1 + b_1 x_1$$
$$y = a_2 + b_2 x_2$$
$$\vdots$$

- But, doesn't provide a way to account for joint effects
- Example: Consider three linear models to predicting longevity:
  - A: Longevity vs. some factor in diet (e.g. amount of fiber consumed)
  - B: Longevity vs. exercise
  - C: Longevity vs. diet AND exercise
  - What does C tell you that A and B do not?

Special Case: Single Variable

- Suppose $k = 1$ predictor.
- Feature matrix and coefficient vector:

$$A = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

- LS soln: $\beta = \left(\frac{1}{N} A^T A\right)^{-1} \left(\frac{1}{N} A^T y\right) = P^{-1} r$

$$P = \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{bmatrix}, \qquad r = \begin{bmatrix} \bar{y} \\ \overline{xy} \end{bmatrix}$$

- Obtain single variable solutions for coefficients (after some algebra):

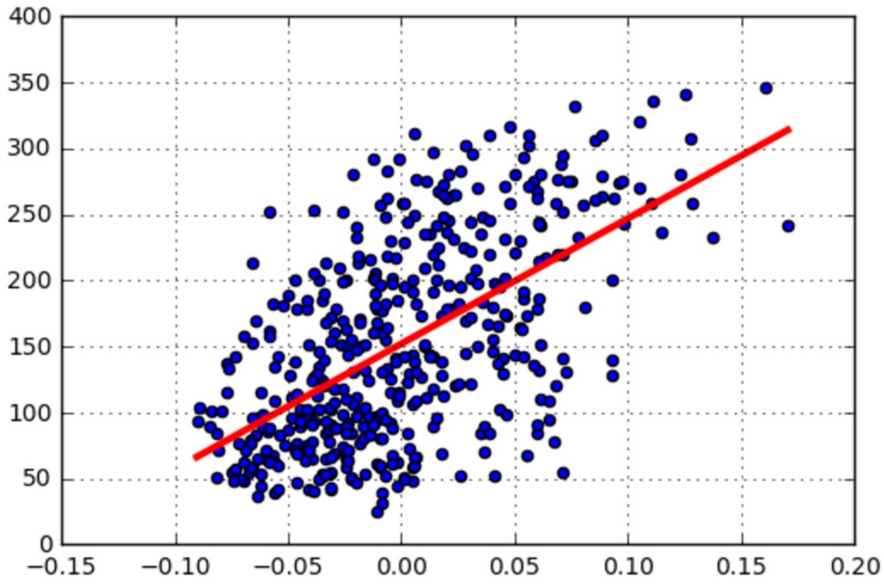
$$\beta_1 = \frac{s_{xy}}{s_x^2}, \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}, \qquad R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}$$

Simple Linear Regression for Diabetes Data

- Try a fit of each variable individually
- Compute R_k^2 coefficient for each variable
- Use formula on previous slide
- “Best” individual variable is a poor fit

Scatter Plot

- No one variable explains glucose well
- Multiple linear regression is much better



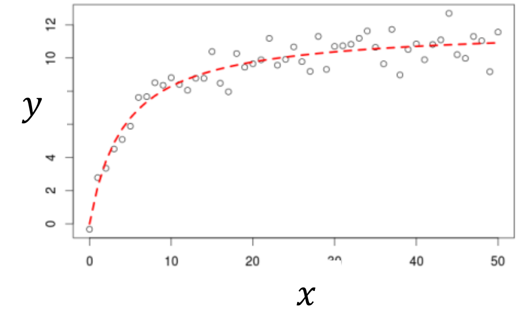
Poorer performance when single variable regression model is used!

Outline

- Motivating Example: Understanding glucose levels in diabetes patients
- Math background
- Multiple variable linear models
- Least squares solutions
- Computing the solutions in python
- Special case: Simple linear regression
- Extensions

Transformed Linear Models

- Standard linear model: $\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$
- Linear in terms of **both** predictors and coefficients
- Linear model may be too restrictive
  - Relation between x and y can be nonlinear



- Useful to look at models in transformed form:

$$\hat{y} = \beta_1 \phi_1(\boldsymbol{x}) + \cdots + \beta_p \phi_p(\boldsymbol{x})$$

  - Each function $\phi_j(\boldsymbol{x}) = \phi_j(x_1, \ldots, x_d)$ is called a basis function
  - Each basis function may be nonlinear and a function of multiple variables
- Can write in vector form: $\hat{y} = \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\beta}$
  - $\boldsymbol{\phi}(\boldsymbol{x}) = [\phi_1(\boldsymbol{x}), \ldots, \phi_p(\boldsymbol{x})]$, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]$

Fitting Transformed Linear Models

- Consider transformed linear model
$$\hat{y} = \beta_1 \phi_1(\boldsymbol{x}) + \cdots + \beta_p \phi_p(\boldsymbol{x})$$
- We can fit this model exactly as before
  - Given data $(\boldsymbol{x}_i, y_i), i = 1, \ldots, N$
  - Want to fit the model from the transformed variables $\phi_j(\boldsymbol{x})$ to target y
  - Define the transformed matrix:

$$A = \begin{bmatrix} \phi_1(\boldsymbol{x}_1) & \cdots & \phi_p(\boldsymbol{x}_1) \\ \vdots & \vdots & \vdots \\ \phi_1(\boldsymbol{x}_N) & \cdots & \phi_p(\boldsymbol{x}_N) \end{bmatrix}$$

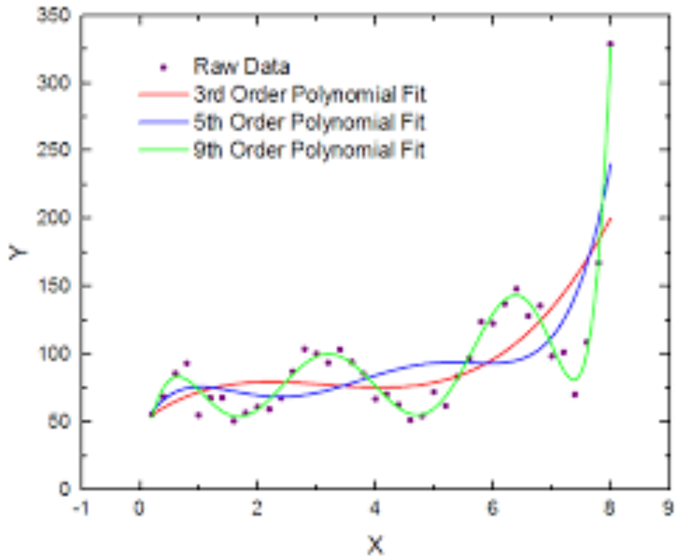
  - Predictions: $\hat{y} = A\beta$
  - Least squares fit $\beta = (A^T A)^{-1} A^T y$

Example: Polynomial Fitting

- Suppose y only depends on a single variable x,
- Want to fit a polynomial model
  - $y \approx \beta_0 + \beta_1 x + \cdots \beta_d x^d$
- Given data $(x_i, y_i), i = 1, \dots, n$
- Take basis functions $\phi_j(x) = x^j, j = 0, \dots, d$
- Transformed model: $\hat{y} = \beta_0\phi_0(x) + \cdots + \beta_d\phi_d(x)$
- Transformed matrix is:

$$\boldsymbol{A} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^d \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & \cdots & x_n^d \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_d \end{bmatrix}$$

  - $p = d + 1$ transformed features from 1 original feature
- Will discuss how to select d in the next lecture



Other Nonlinear Examples

- Multinomial model: $\hat{y} = a + b_1 x_1 + b_2 x_2 + c_1 x_1^2 + c_2 x_1 x_2 + c_3 x_2^2$
  - Contains all second order terms
  - Define parameter vector $\beta = [a, b_1, b_2, c_1, c_2, c_3]$
  - Transformed vector $\phi(x_1, x_2) = [1, x_1, x_2, x_1^2, x_1 x_2, x_2^2]$
  - Note that the features are nonlinear functions of $\boldsymbol{x} = [x_1, x_2]$

- Exponential model: $\hat{y} = a_1 \mathrm{e}^{-b_1 x} + \cdots + a_d e^{-b_d x}$
  - If the parameters $b_1, \ldots, b_d$ are fixed, then the model is linear in the parameters $a_1, \ldots, a_d$
  - Parameter vector $\beta = [a_1, \ldots, a_d]$
  - Transformed vector $\phi(x) = [e^{-b_1 x}, \ldots, e^{-b_d x}]$
  - But, if the parameters $b_1, \ldots, b_d$ are not fixed, the model is nonlinear in $b_1, \ldots, b_d$

Linear Models via Re-Parametrization

- Sometimes models can be made into a linear model via re-parametrization
- Example: Consider the model: $\hat{y} = Ax_1(1 + Be^{-x_2})$
  - Meaning you want to use this model, but how to find optimal A and B?
  - Parameters (A, B)
- This is nonlinear in (A, B) due to the product AB: $\hat{y} = Ax_1 + ABx_1e^{-x_2}$
- But, we can define a new set of parameters:
  - $\beta_1 = A$ and $\beta_2 = AB$
- Then, $\hat{y} = \beta_1 x_1 + \beta_2 x_1 e^{-x_2}$
- Basis functions: $\phi(x_1, x_2) = [x_1, x_1 e^{-x_2}]$
- After we solve for β_1, β_2 we can recover A, B via inverting the equations:

$$A = \beta_1, \qquad B = \frac{\beta_2}{A}$$

One Hot Coding

Model	ϕ_1	ϕ_2	ϕ_3
Ford	1	0	0
BMW	0	1	0
GM	0	0	1

- Suppose that one feature x_j is a categorical variable
- Ex: Predict the price of a car, y, given model x_1 and interior space x_2
  - Suppose there are 3 different models of a car (Ford, BMW, GM)
  - Bad idea: Arbitrarily assign an index to each possible car model
  - Can give unreasonable relations
- One-hot coding example:
  - With 3 possible categories, represent x_1 using 3 binary features (ϕ_1, ϕ_2, ϕ_3)
  - Model: $y = \beta_0 + \beta_1\phi_1 + \beta_2\phi_2 + \beta_3\phi_3 + \beta_4 x_2$
  - Essentially obtain 3 different models:
    - Ford: $y = \beta_0 + \beta_1 + \beta_4 x_2$
    - BMW: $y = \beta_0 + \beta_2 + \beta_4 x_2$
    - GM: $y = \beta_0 + \beta_3 + \beta_4 x_2$
  - Allows different intercepts (or mean values) for different categories!