Machine Learning

CSE 142

Xin (Eric) Wang

Friday, October 22, 2021

Today

- Linear learning models, Ch. 7

Notes

- Midterm exam – Monday, November 1st (in class)
  - Material covered: Everything through next Wednsday
    - Lectures, reading, discussion sessions, homeworks
    - No Python questions
  - Virtual in-class exam
    - With camera on all the time (all the teaching staffs will be proctoring);
    - No phones; No earphones;
    - No Google search; No keyboard typing;
    - Write answers on a white paper (or iPad) using your pen;
    - Picture and upload it to Canvas/Gradescope before the end time;
    - I’ll also provide some information, formulas, etc.
  - Brief review in class next week
  - A practice midterm will be posted next week (including provided info/formulae that will be on the midterm)

Key statistical concepts

- Mean – average; expected value of a variable

  $\mu_x = E[X] = \sum_{i=1}^{n} x_i p_i$ or $\int x\, p(x)\, dx$

- Variance – a measure of the spread of a variable

  $Var(X) = \sigma_x{}^2 = E[(X - \mu_x)^2] = E[X^2] - \mu_x{}^2$

  Standard deviation: $\sigma_x = \mathrm{Sqrt}(\sigma_x{}^2)$

- Estimating mean and variance from data $\{x_i\}$

  Sample mean: $\hat{\mu}_x = \frac{1}{n}\sum_i x_i$

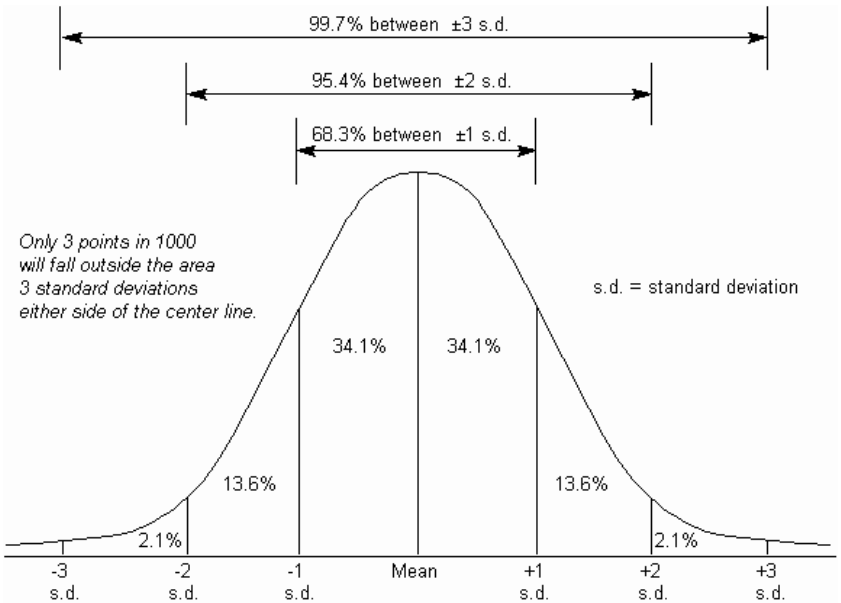
  Sample variance: $\hat{\sigma}_x{}^2 = \frac{1}{n}\sum_i (x_i - \hat{\mu}_x)^2$ or $s = \frac{1}{n-1}\sum_i (x_i - \hat{\mu}_x)^2$

- Covariance – a measure of how two variables change together

  $Cov(X,Y) = \sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)] = E[XY] - \mu_x \mu_y$

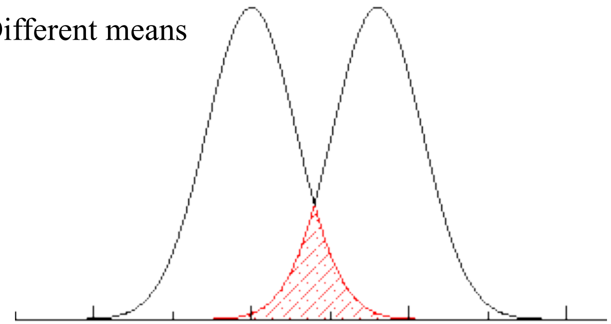
  Sample covariance: $\hat{\sigma}_{xy} = \frac{1}{n}\sum_i (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)$ or $\frac{1}{n-1}\sum_i (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)$

Key statistical concepts (cont.)

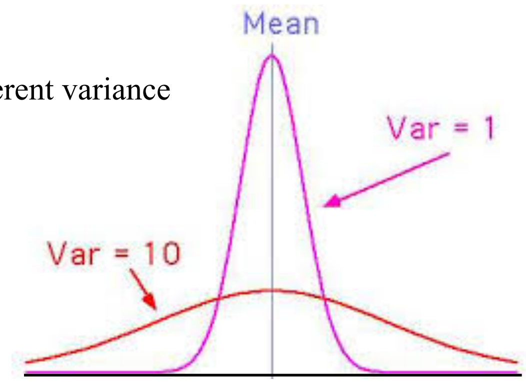


Gaussian (normal) distribution

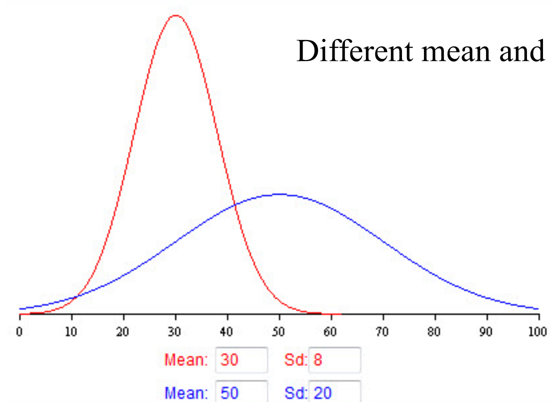
Different means



Different variance



Different mean and variance

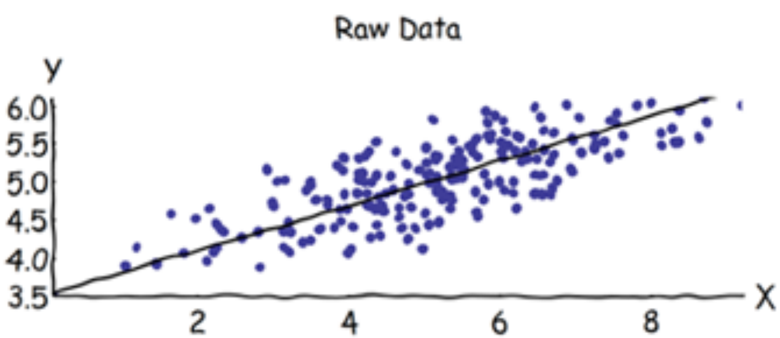


Key statistical concepts (cont.)

- Covariance matrix Σ
  - For n variables $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$, Σ is an n x n matrix whose elements are $\mathrm{Cov}(X_i, X_j)$
  - Diagonal entries are variances: $Cov(X_i, X_i) = Var(X_i)$
- If variables x and y are uncorrelated, then

$$Cov(X, Y) = \sigma_{xy} = 0$$

  - Uncorrelated variables: knowing the value of X (or Y) tells you nothing about the value of Y (or X)
  - So the covariance matrix for uncorrelated variables is a diagonal matrix consisting of the n variances
- If $Cov(X, Y) > 0$, then Y tends to increase as X increases



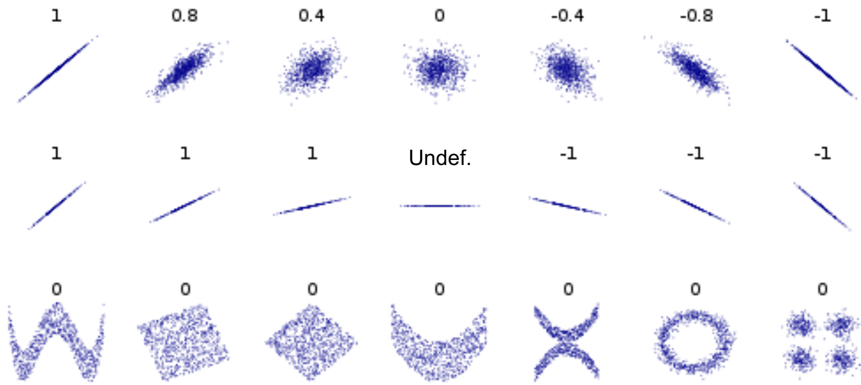
Non-zero (positive) covariance

Examples

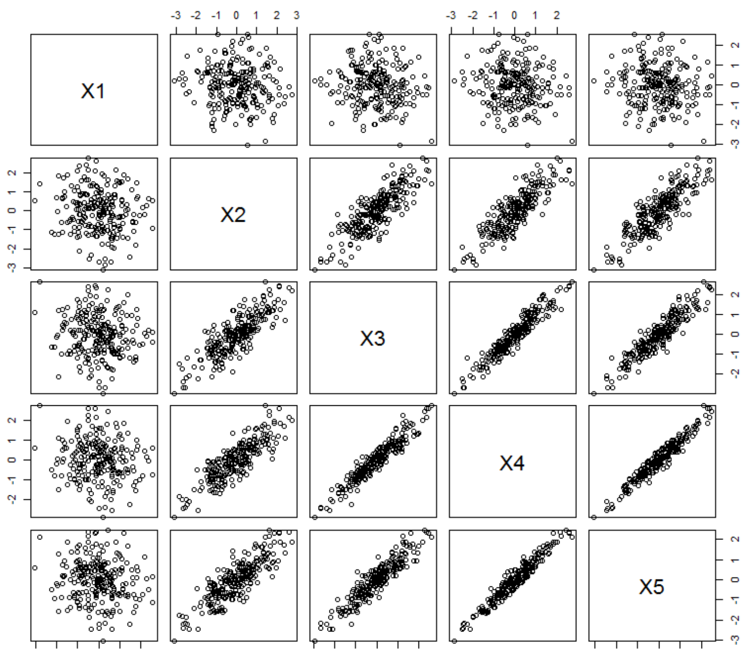
2D data and their correlation coefficient (ρ) values

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

$$-1 \leq \rho \leq 1$$



Not a useful measure for nonlinear data!



Visualizing a 5-variable covariance matrix (symmetric about the diagonal)

Linear models

- Linear models are geometric models for which the regression functions or decision boundaries are linear
  - Lines, planes, hyperplanes (N-dimensional planes)

- Definition of a linear function:

$$y = f(ax_1 + bx_2) = af(x_1) + bf(x_2)$$

or in matrix notation, a linear transformation:

$$\boldsymbol{y} = M\boldsymbol{x}$$

- An affine function is a linear function plus a constant

$$f_{\text{aff}}(x) = f_{\text{lin}}(x) + c$$

In matrix notation:

$$\boldsymbol{y} = M\boldsymbol{x} + \boldsymbol{c}$$

Using homogeneous coordinates:

$$\boldsymbol{y} = \boldsymbol{M}'\boldsymbol{x}_{\boldsymbol{h}}$$

$$\boldsymbol{y} = M\boldsymbol{x} + \boldsymbol{c}$$

$$\boldsymbol{y} = [\boldsymbol{M} \quad \boldsymbol{c}] \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix}$$

$$\boldsymbol{y} = M'\boldsymbol{x}_{\boldsymbol{h}}$$

$$\boldsymbol{y} = M\boldsymbol{x} + \boldsymbol{c}$$

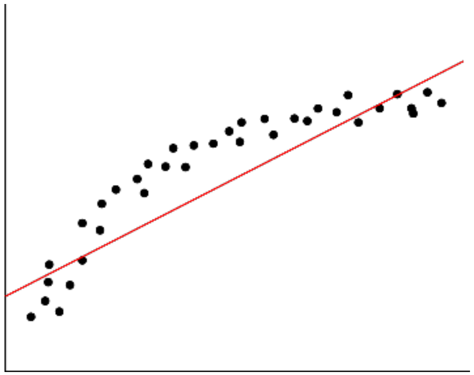
$$\begin{bmatrix} \boldsymbol{y} \\ 1 \end{bmatrix} = \begin{bmatrix} M & \boldsymbol{c} \\ \boldsymbol{0} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix}$$

$$\boldsymbol{y}_{\boldsymbol{h}} = M''\boldsymbol{x}_{\boldsymbol{h}}$$

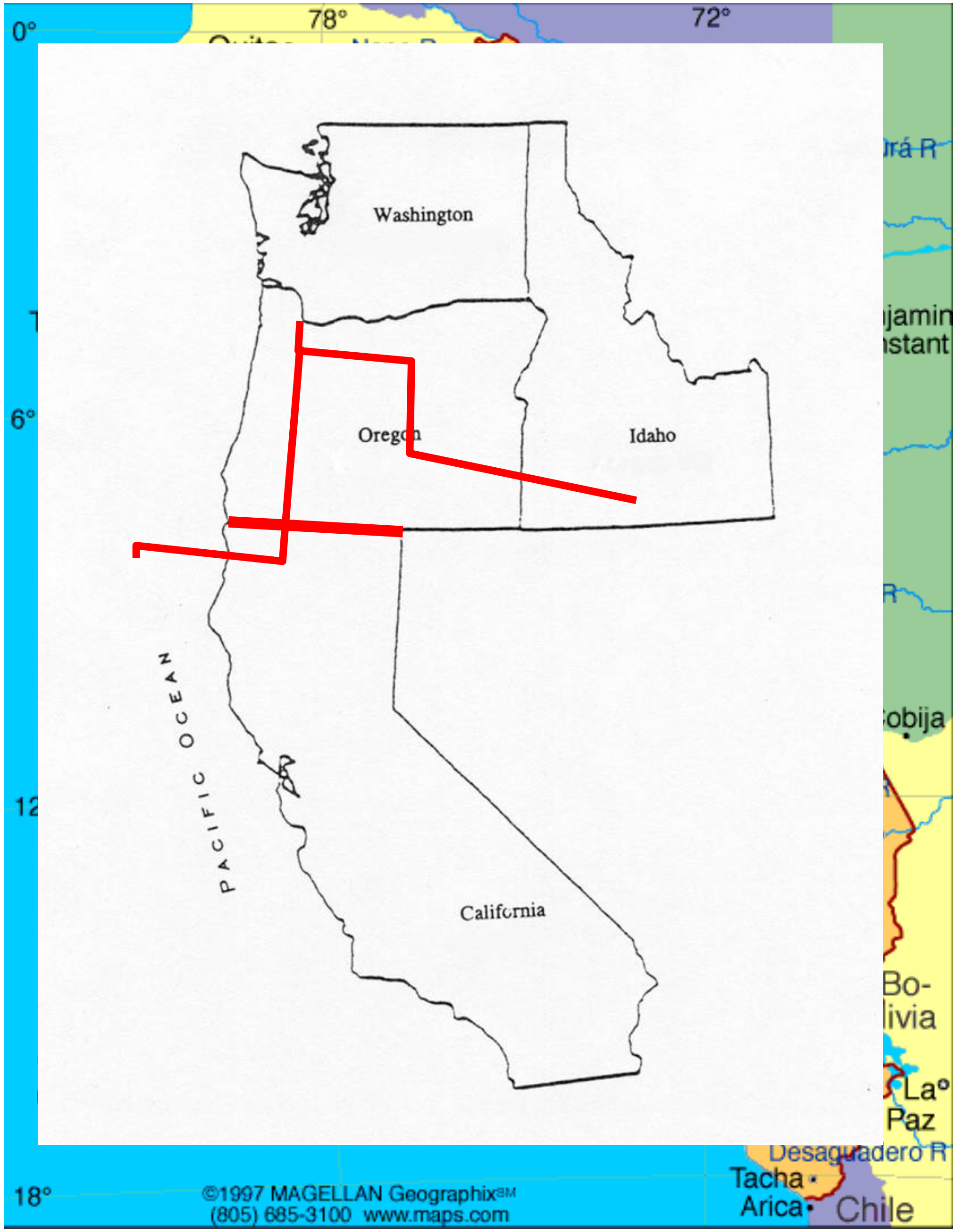
So we can use the term ***linear models*** to include ***affine models***

Linear models

- Linear learning models are widely used because
  - Many functions can be reasonably approximated as linear, or at least as piecewise linear
  - They're simple, and thus easy to train
  - The math is tractable
  - They avoid over-fitting – i.e., they generalize well when the data is very noisy
- However, they are prone to under-fitting
  - I.e., over-simplifying a more complicated function



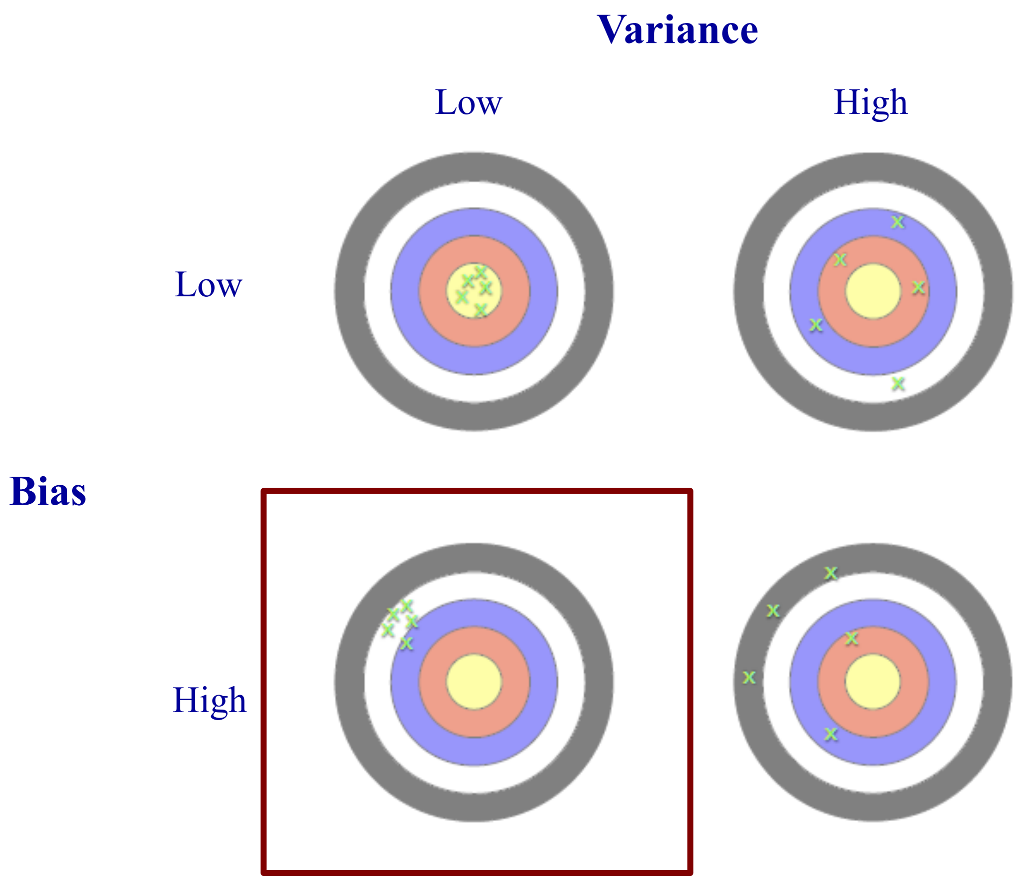
- For example, learning borders from sample data



The border between California and Oregon – linear
The border between Texas and New Mexico – piecewise linear
The border between Texas and Oklahoma – piecewise linear approx.
The border between Peru and Brazil – complicated!

Linear models

- Linear models tend to have low variance but high bias



Low variance – stability, robustness

Performance on different testing sets should be similar

High bias – limited accuracy, underfitting, systematic (but consistent) errors

Parametric models

- Linear models are parametric models
  - Within a given family of models (e.g., lines or planes), we just need to learn a small number of model parameters (e.g., 2 or 3 coefficients)
- We'll also consider nonparametric models
  - No explicit assumption about the shape of the model (the form of the mapping function)
- For example, in a 2D classification problem we could use linear decision boundaries (lines) as a parametric model, or the nearest-neighbor approach (minimum distance) as a non-parametric model
- This distinction is also important in density estimation – estimating a probability distribution or density from data
  - E.g., in parametric estimation, we might assume the pdf is Gaussian, so the task becomes estimating the Gaussian parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Linear least-squares regression

- Regression learns a function (the regressor) that is a mapping $\hat{f} : \mathcal{X} \rightarrow \mathbb{R}$; it's learned from examples, $(x_i, f(x_i))$
  - I.e., the target variable (output $\hat{f}(x)$) is real-valued
- Linear regression – the function is linear
  - Fit a line/plane/hyperplane to the data
- The difference between f and $\hat{f}$ is known as the residual ϵ

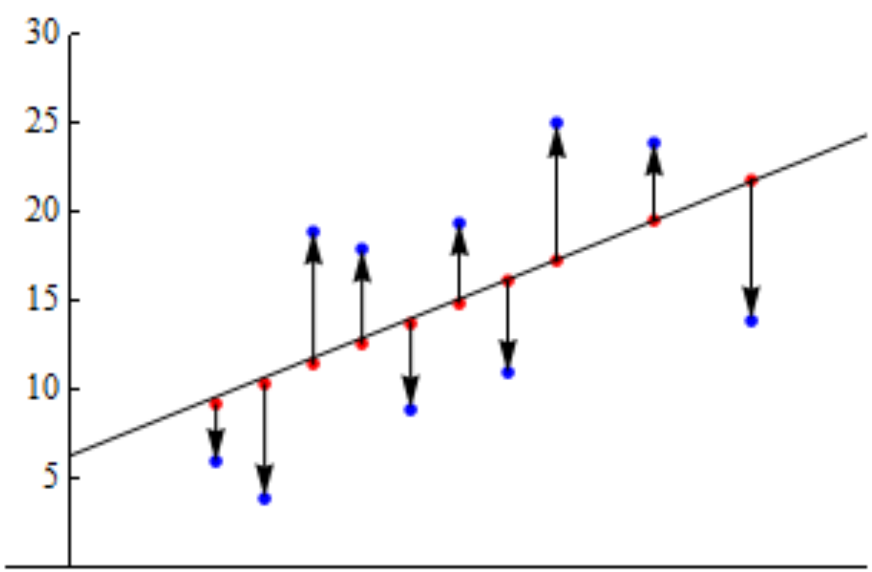
$$\epsilon_i = f(x_i) - \hat{f}(x_i)$$

- The least squares method minimizes the sum of the squared residuals – i.e., find $\hat{f}$ that minimizes $\sum_i {\epsilon_i}^2$ on the training data
- Univariate or multivariate regression
  - Univariate – one input variable
  - Multivariate – multiple input variables

Note: In Statistics, multivariate regression means multiple targets (outputs). ML sources may use the term incorrectly, but I'll stick here with the book's usage, where multivariate means multiple input variables.

Linear least-squares regression example

- We wish to find the relationship between the height and weight of adults
  - Data: n measurements, $(h_i, w_i) \rightarrow$ (*input*, *output*)
  - Parametric linear model: $w = a + bh \quad \Rightarrow \quad w_i = a + bh_i + \epsilon_i$
  - Residual: $\epsilon_i = w_i - (a + bh_i)$
  - Find (a, b) that minimizes $\sum_i [w_i - (a + bh_i)]^2$ on the training data



Linear least-squares regression example

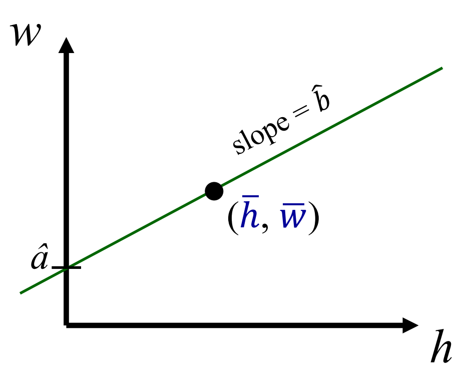
- To minimize $\sum_i [w_i - (a + bh_i)]^2$, set the partial derivatives (wrt a and b) to zero and solve for a and b

$$\frac{\partial}{\partial a} \sum_{i=1}^{n} (w_i - (a + bh_i))^2 = -2 \sum_{i=1}^{n} (w_i - (a + bh_i)) = 0 \qquad \Rightarrow \hat{a} = \overline{w} - \hat{b}\overline{h}$$

$$\frac{\partial}{\partial b} \sum_{i=1}^{n} (w_i - (a + bh_i))^2 = -2 \sum_{i=1}^{n} (w_i - (a + bh_i))h_i = 0 \qquad \Rightarrow \hat{b} = \frac{\sum_{i=1}^{n} (h_i - \overline{h})(w_i - \overline{w})}{\sum_{i=1}^{n} (h_i - \overline{h})^2}$$

- So the regression model is $\boxed{w = \hat{a} + \hat{b}h = \overline{w} + \hat{b}(h - \overline{h})}$

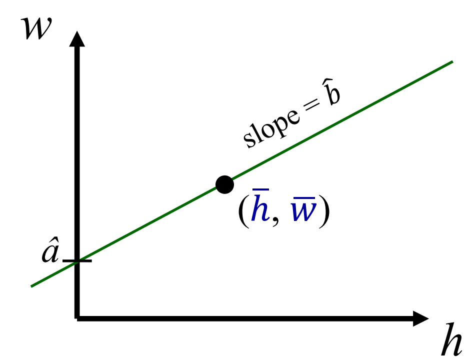
Note that the regression line goes through $(\bar{h}, \overline{w})$



The regression coefficient

- The slope ($\hat{b}$) is the regression coefficient

$$\hat{b} = \frac{\sum_{i=1}^{n}(h_i - \overline{h})(w_i - \overline{w})}{\sum_{i=1}^{n}(h_i - \overline{h})^2} = \frac{n\sigma_{hw}}{n{\sigma_h}^2} = \frac{\sigma_{hw}}{{\sigma_h}^2}$$



- In general, the regression coefficient for a feature x and a target variable y is

$$\hat{b} = \frac{\overbrace{\sigma_{xy}}^{\text{covariance}(x,\,y)}}{\underbrace{{\sigma_x}^2}_{\text{variance}(x)}}$$

- We often simplify the problem by first normalizing the data
  - Find the data averages $(\overline{h}, \overline{w})$
  - Subtract the averages from the data: $h_i \leftarrow h_i - \overline{h}$
    $w_i \leftarrow w_i - \overline{w}$

- This makes $\hat{a} = 0$, so we're just left with estimating the regression coefficient $\hat{b}$

Quiz: Tesla Stock Prediction

Suppose we have three data points of the Tesla stock prices: \$69 in Year 1, \$123 in Year 2, and \$168 in Year 3. Can you predict its stock price in Year 4 using linear regression?

$$w = \hat{a} + \hat{b}h = \overline{w} + \hat{b}(h - \overline{h})$$

$$\hat{b} = \frac{\sum_{i=1}^{n}(h_i - \overline{h})(w_i - \overline{w})}{\sum_{i=1}^{n}(h_i - \overline{h})^2} = \frac{n\sigma_{hw}}{n{\sigma_h}^2} = \frac{\sigma_{hw}}{{\sigma_h}^2}$$

Multivariate linear regression

- Most linear regression problems involve multiple input variables
  - E.g., estimate a patient's cholesterol level from several input variables
- In multivariate LR, there are N+1 regression parameters
- Linear regression equations:

x_{i0} = 1 (homogeneous notation)

Univariate

$$y_i = w_1 x_i + w_0 + \epsilon_i$$

Multivariate

$$y_i = w_2 x_{i2} + w_1 x_{i1} + w_0 x_{i0} + \epsilon_i$$

Column of 1s

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} x_{12} & x_{11} & \boxed{x_{10}} \\ x_{22} & x_{21} & x_{20} \\ \vdots & \vdots & \vdots \end{bmatrix} \qquad \boldsymbol{w} = \begin{bmatrix} w_2 \\ w_1 \\ w_0 \end{bmatrix} \qquad \boldsymbol{\epsilon_i} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \end{bmatrix}$$

Labels | Data (homogeneous) | Regression parameters | Residuals

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$$