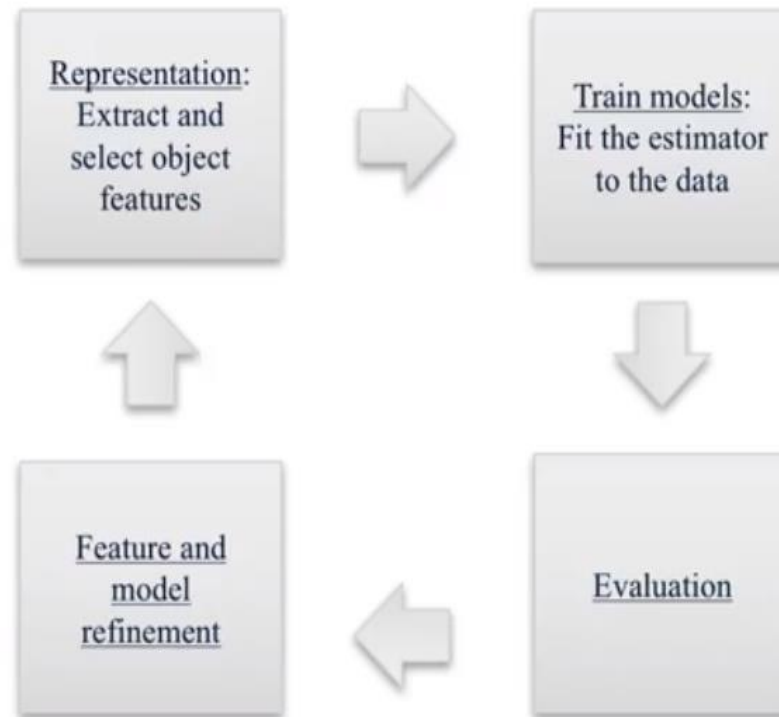**COMP9321:**
Data services engineering

Week 8: Linear Regression

Term 1, 2021
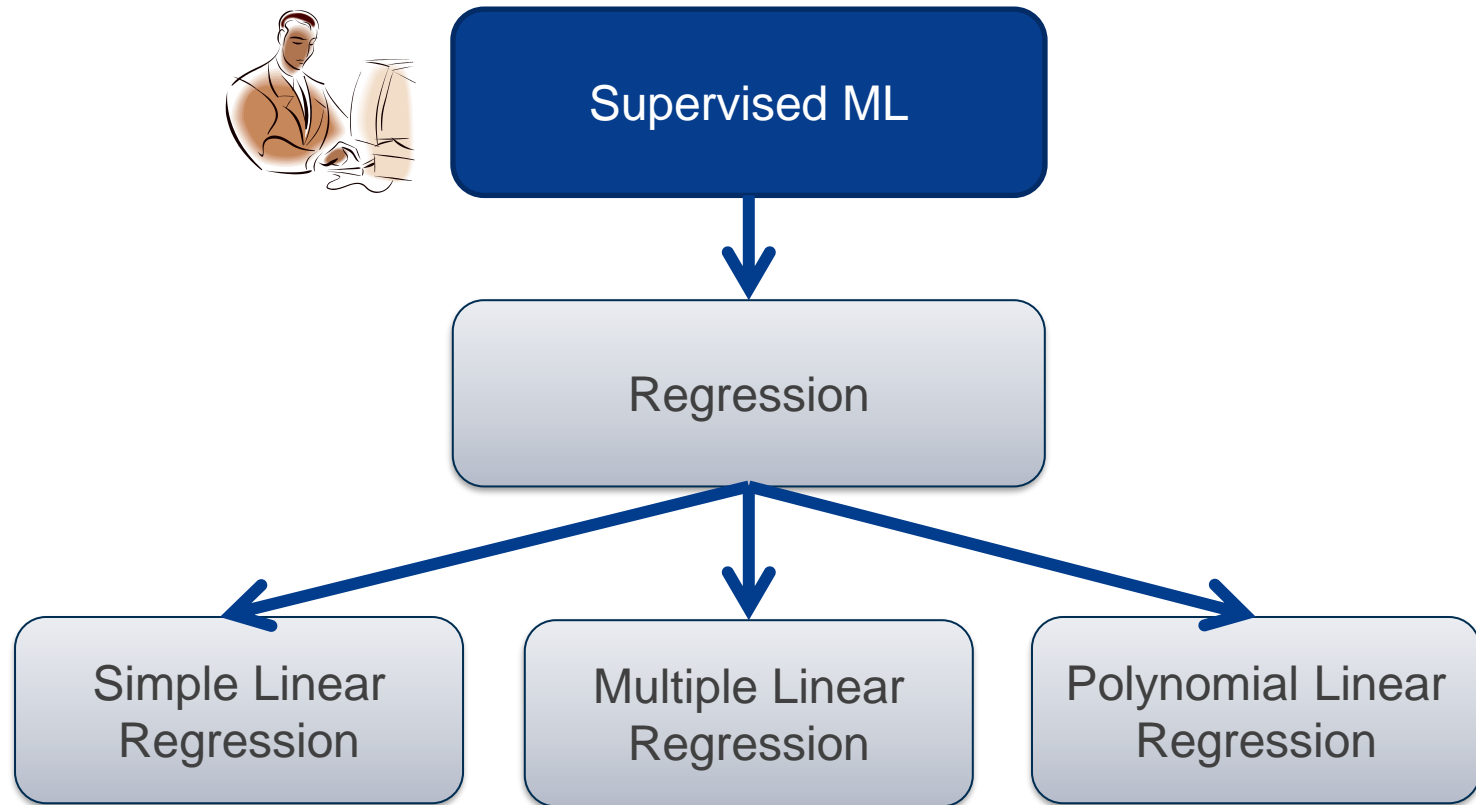
By Mortada Al-Banna, CSE UNSW

Refresher

Represent / Train / Evaluate / Refine Cycle

Regression Analysis

Supervised ML

Regression

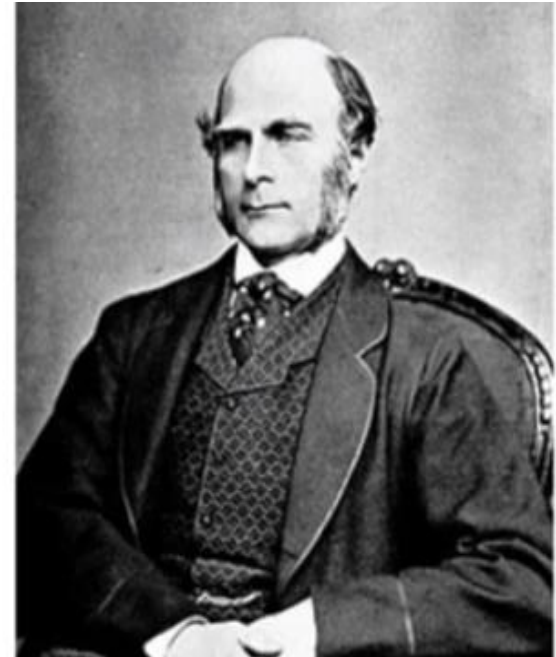Simple Linear Regression   Multiple Linear Regression   Polynomial Linear Regression
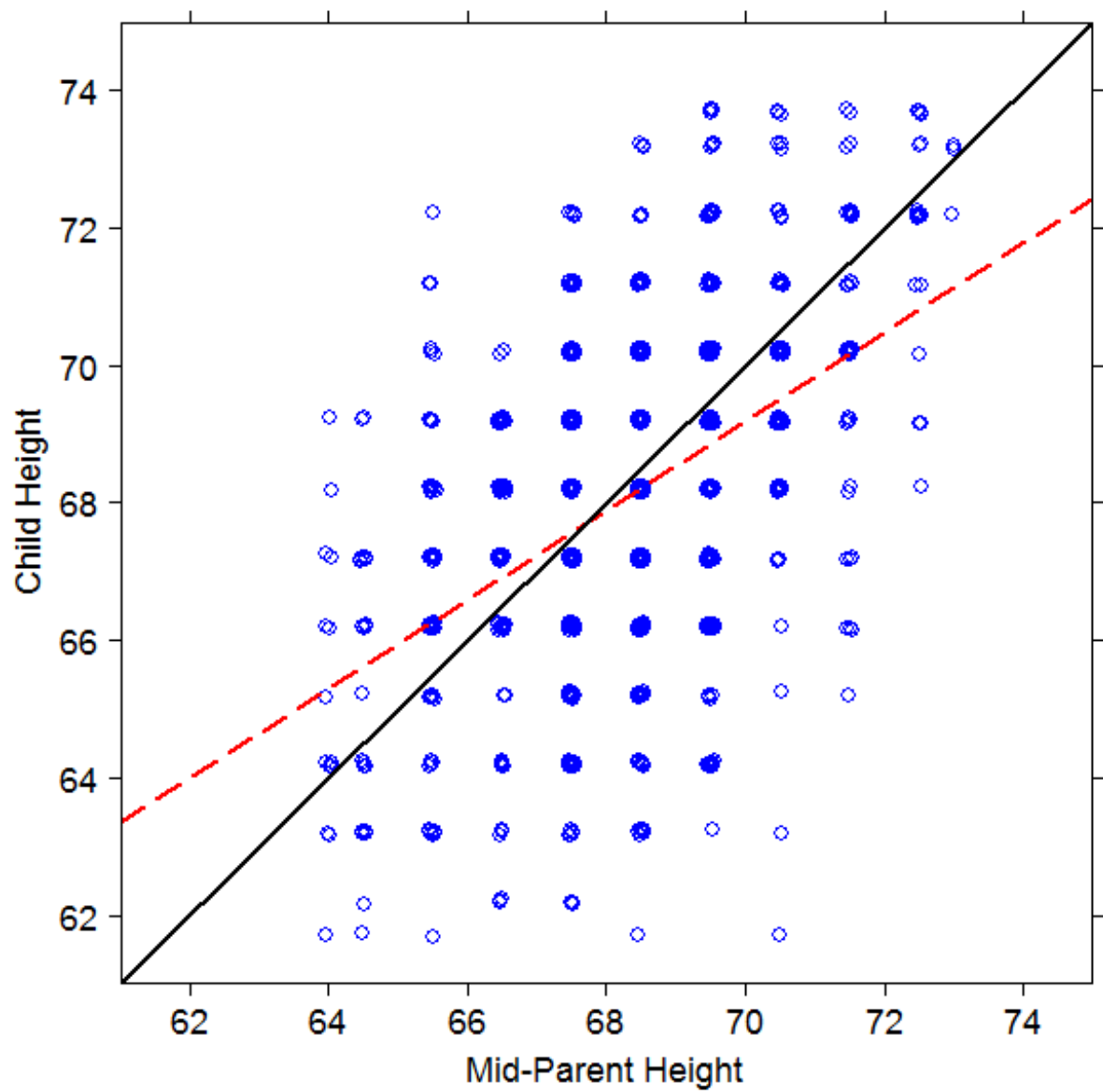
Sir Francis Galton, 1822-1911

Regression Towards Mediocrity in
Hereditary Stature

> *Journal of the Anthropological
> Institute*, 1886; 15:246-63

Regression Analysis

- A linear Model is a sum of weighted variables that predict a target output value given an input data instance

Example: Predicting housing prices

House features: taxes paid per year (X_{tax}), age in years (X_{age})

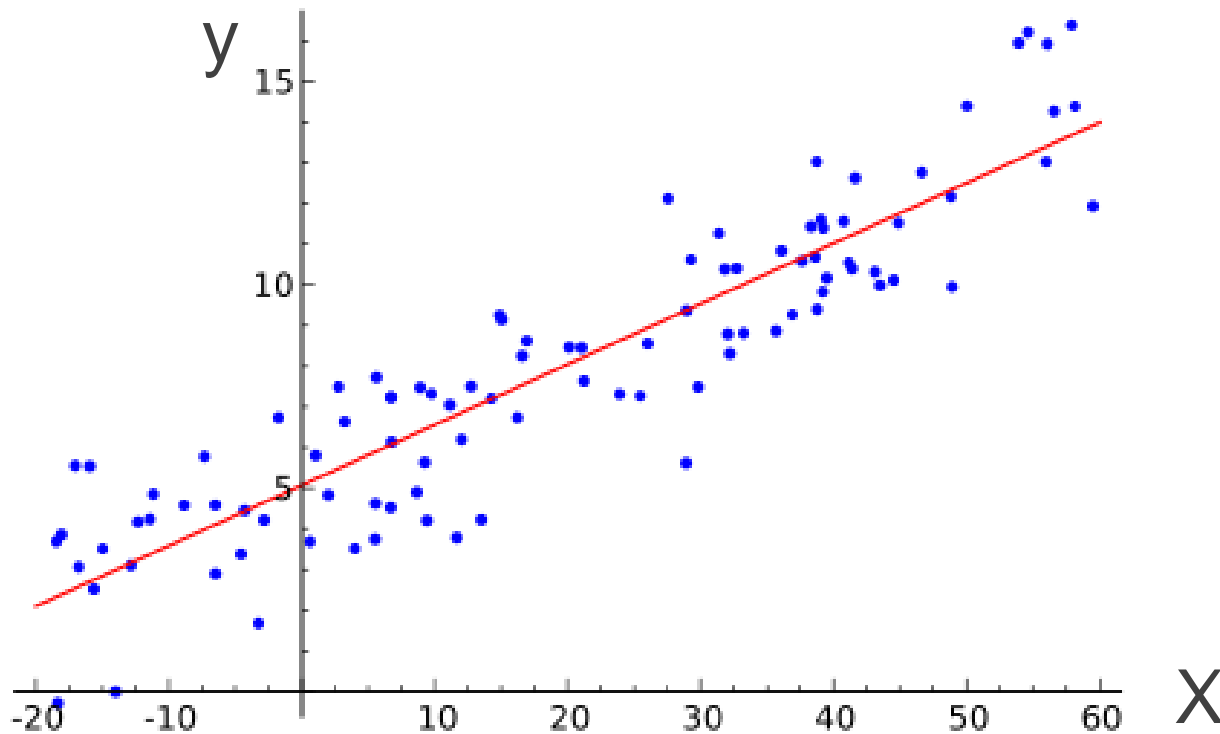*Predicted price*$= 80000+ 100\ X_{tax} - 4000\ X_{age}$

- So if the house tax per year is 20000, and the age of the house is 60 years then the predicted selling price is:

Predicted price$=80000+100 \times 20000 - 4000 \times 60 = 1,840,000$

Linear Regression

We want to find the "best" line (linear function
y=f(X)) to explain the data.

Linear Regression

The predicted value of y is given by:

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j$$

The vector of coefficients $\hat{\beta}$ is the regression model.

Linear Regression

The regression formula $\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j + \varepsilon$

e.g., j = 1   Random error

$$\widehat{y} = \hat{\beta}_0 + X_1 \hat{\beta}_1 + \varepsilon$$

            predictor   Slope of the line

Intercept (where the line crosses y-axis)

The slope and intercept of the line are called regression
   coefficients, model parameters

Our goal is to estimate the model parameters

Assumptions When using Linear Regression

- Outcome Variable must be continuous.
- Linear Relationship between the features and target.
- Little or no Multicollinearity between the features.
- Normal distribution of error terms.
- Minimum Outliers

How to check these assumptions?

- Plotting (e.g., scatter plot, Histogram)
- Calculating coefficients

Correlation Coefficient

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[\, n\sum x^2 - (\sum x)^2 \,][\, n\sum y^2 - (\sum y)^2 \,]}}$$

If r

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

UNSW
SYDNEY

Linear Regression

The regression formula $\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j + \varepsilon$

e.g., j = 1   Random error

$$\hat{y} = \hat{\beta}_0 + X_1 \hat{\beta}_1 + \varepsilon$$

            predictor   Slope of the line

Intercept (where the line crosses y-axis)

The slope and intercept of the line are called regression coefficients, model parameters

Our goal is to estimate the model parameters

Challenge

- Find the values of β_0 and β_1 that the line corresponding to those values is the best fitting line or gives the minimum error (minimum cost)

- Possible solution is to use the Least Square Error solution

- But where do we start and how we determine the proposed line? Gradient descent

Least Square Error Solution

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n$$

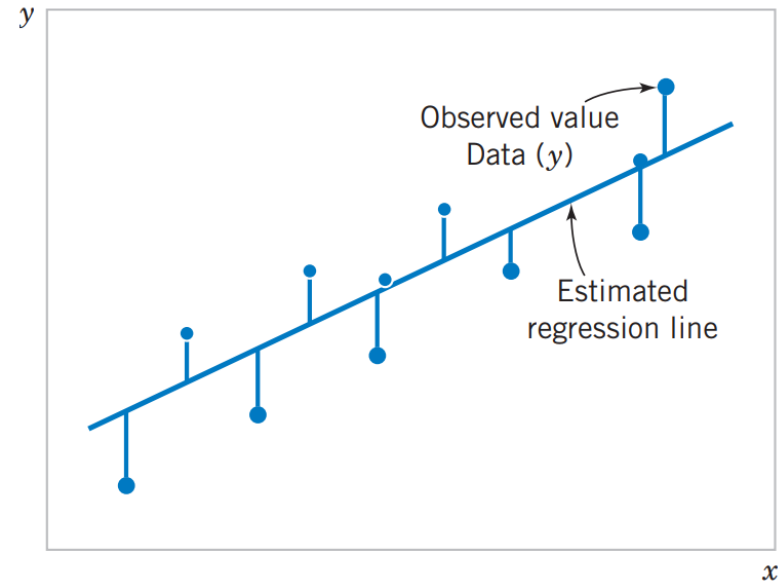To estimate (β_0, β_1) , we find values that minimize squared error

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Solution:

Observed value
Data (y)

Estimated
regression line

$$\left. \frac{\partial L}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$
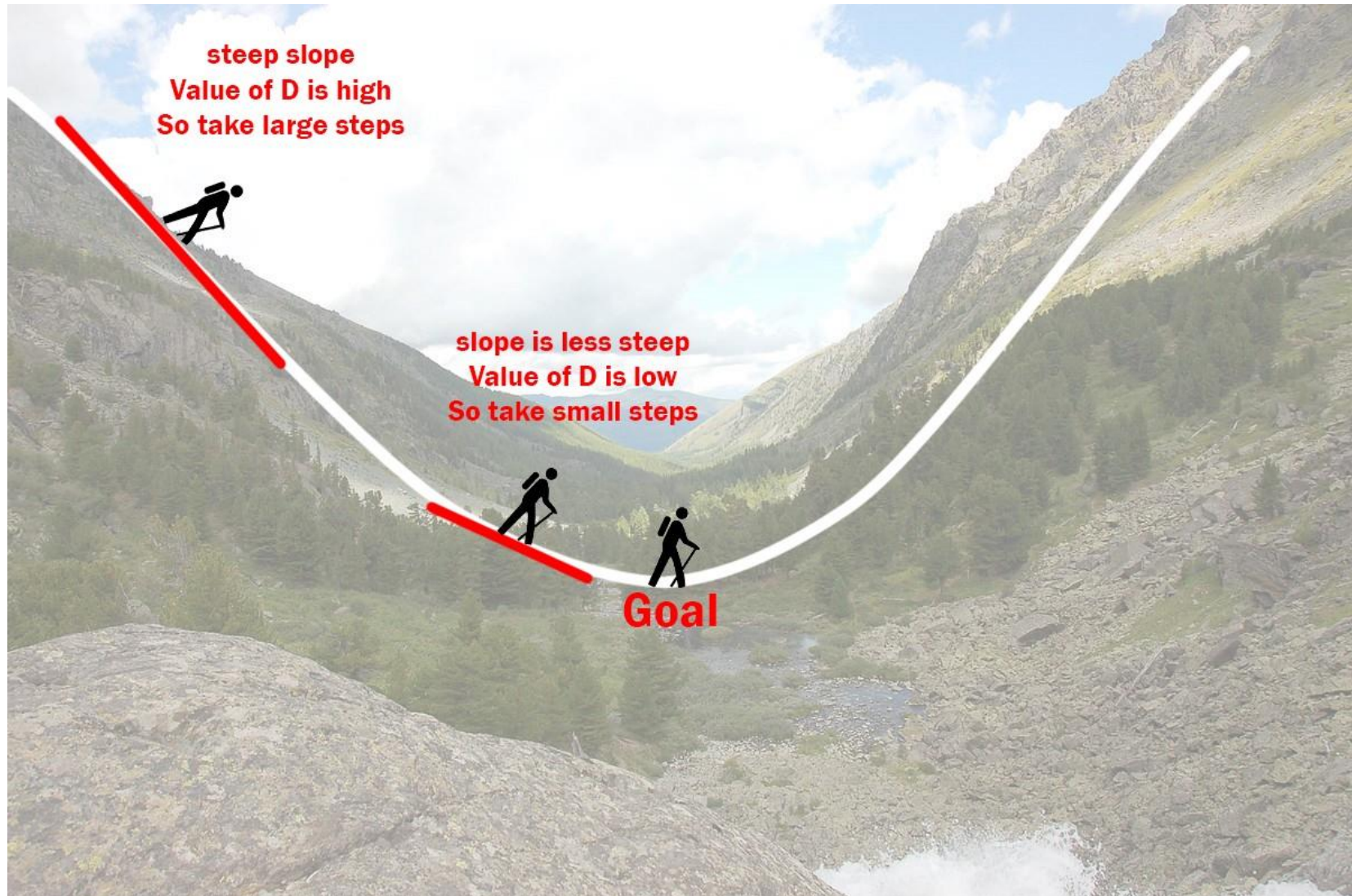
$$\left. \frac{\partial L}{\partial \beta_1} \right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

What is Gradient Descent?

Gradient Descent

- Gradient descent is a method of updating β_0 and β_1 to reduce the cost function(Least Square Error).

- The idea is that we start with some values for β_0 and β_1 and then we change these values iteratively to reduce the cost. Gradient descent helps us on how to change the values.

- To update β_0 and β_1 , we take gradients from the cost function. To find these gradients, we take partial derivatives with respect to β_0 and β_1.

- A smaller learning rate could get you closer to the minima but takes more time to reach the minima, a larger learning rate converges sooner but there is a chance that you could overshoot the minima.
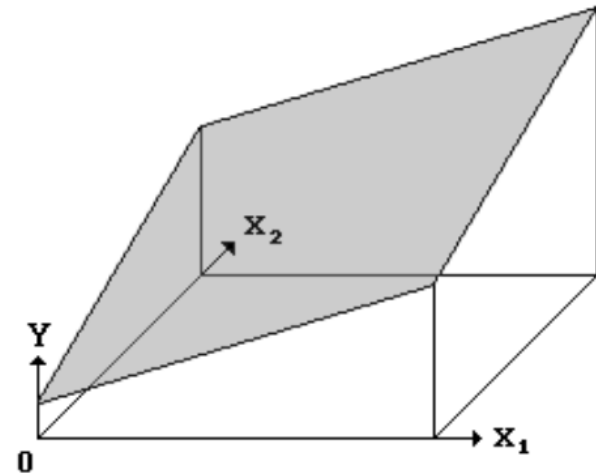
Linear Regression

Simple linear regression
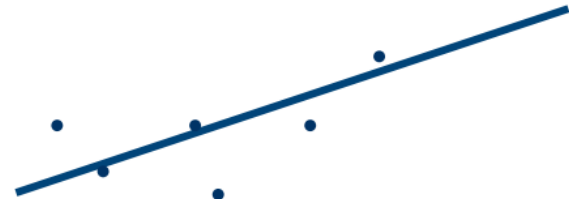
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Multiple Linear Regression Vs Simple Linear Regression

- Simple Linear regression compares the response of a dependent variable given a change in some explanatory variable.

- However, it is **rare** that a dependent variable is explained by **only one variable**. In this case, we use multiple regression, which attempts to explain a dependent variable using **more than one independent variable**.

- Multiple regressions are based on the assumption that there is a **linear relationship** between both the dependent and independent variables. It also assumes **no major correlation** between the **independent variables**.

Multiple Linear Regression (Example)

- Predicting Exxon Mobil (XOM) stock price
  - Option1: Simple Linear Regression rely on the value of the S&P 500 index as the independent variable, or predictor, and the price of XOM as the dependent variable. **Problem**: Unrealistic and provide lower accuracy.
  - Option2: Multiple Linear Regression depends on more than just the performance of the overall market. Other predictors such as the price of oil, interest rates, and the price movement of oil futures can affect the price of XOM and stock prices of other oil companies

UNSW
SYDNEY

Multiple Linear Regression (Example) Cont'd

- Examines how multiple independent variables are related to one dependent variable.

- Once each of the independent factors has been determined to predict the dependent variable, the information on the multiple variables can be used to create an accurate prediction on the level of effect they have on the outcome variable.

Multiple Linear Regression (Example) Cont'd

y_i = dependent variable: price of XOM

x_{i1} = interest rates

x_{i2} = oil price

x_{i3} = value of S&P 500 index

x_{i4} = price of oil futures

β_0 = y-intercept at time zero

β_1 = regression coefficient that measures a unit change in the dependent variable when x_{i1} changes - the change in XOM price when interest rates change

β_2 = coefficient value that measures a unit change in the dependent variable when x_{i2} changes – the change in XOM price when oil prices change

β_3 = coefficient value that measures a unit change in the dependent variable when x_{i3} changes – the change in XOM price when the value of S&P 500 index change

β_4 = coefficient value that measures a unit change in the dependent variable when x_{i4} changes – the change in XOM price when price of oil futures change

$$Y_i = \beta 0 + \beta 1 X1 + \beta 2 X2 + \beta 3 X3 + \beta 4 X4 + \varepsilon$$

UNSW
SYDNEY

Multiple Linear Regression (Selecting the Features)

- The multiple linear regression explains the relationship between **one continuous dependent variable** (*y*) and **two or more independent variables** (β1, β2, β3… etc.)

- Challenge: How to determine which features to keep and which to toss?

  ➢ **Chuck Everything In and Hope for the Best**

  ➢ **Backward Elimination**

  ➢ **Forward Selection**

  ➢ **Bidirectional Elimination**

Something You Need to Know (P-Value)

- The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable.

- Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

- Null Hypothesis: It's like if you were doing a trial of a drug that doesn't work. In that trial, there just wouldn't be a difference between the group that took the drug and the rest of the population. The difference would be null. You always assume that the null hypothesis is true until you have evidence that it isn't.

UNSW
SYDNEY

Backward Elimination

1. First, you'll need to set a significance level for which data will stay in the model. For example, you might want to set a significance level of 5% (SL = 0.05). This is important and can have real ramifications, so give it some thought.
2. Next, you'll fit the full model with all possible predictors.
3. You'll consider the predictor with the highest **P-value**. If your **P-value** is greater than your significance level, you'll move to step four, otherwise, you're done!

UNSW
SYDNEY

Backward Elimination

4. Remove that predictor with the highest P-value.
5. Fit the model without that predictor variable. If you just remove the variable, you need to refit and rebuild the model. The coefficients and constants will be different. When you remove one, it affects the others.
6. Go back to step 3, do it all over, and keep doing that until you come to a point where even the highest P-value is < SL. Now your model is ready. All of the variables that are left are less than the significance level.

Correlation and Collinearity

- Checking for collinearity helps you get rid of variables that are skewing your data by having a **significant relationship** with another variable

- **Correlation** between variables describe the **relationship** between two variables. If they are **extremely correlated**, then they are **collinear**

- Having high collinearity (correlation of 1.00) between predictors will affect your coefficients and the accuracy, plus its ability to reduce the LSE (Least Squared Errors)

- The simplest method to detect collinearity would be to plot it out in graphs or to view a correlation matrix to check out pairwise correlation.

Useful Resources

- https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a

- https://medium.com/@powersteh/an-introduction-to-applied-machine-learning-with-multiple-linear-regression-and-python-925c1d97a02b

- https://machinelearningmastery.com/linear-regression-for-machine-learning/

- https://www.investopedia.com/terms/m/mlr.asp

- https://towardsdatascience.com/multiple-linear-regression-in-four-lines-of-code-b8ba26192e84

Questions?