Linear regression and logistic regression

Machine Learning Summer Course 2020

Krishnakant Saboo

27th June 2020

What it takes to fly….

An airline firm is starting flights on a new route of 8000km. Since they have not flown flights at this distance before, they are unaware of amount of fuel needed. Fuel is expensive so it is important to as closely estimate the amount of fuel needed for the journey as possible. The airline firm has historical data from its other flights. Can they leverage that to find the fuel requirement? If yes, then how?

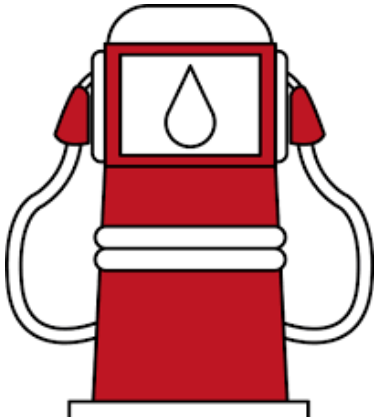
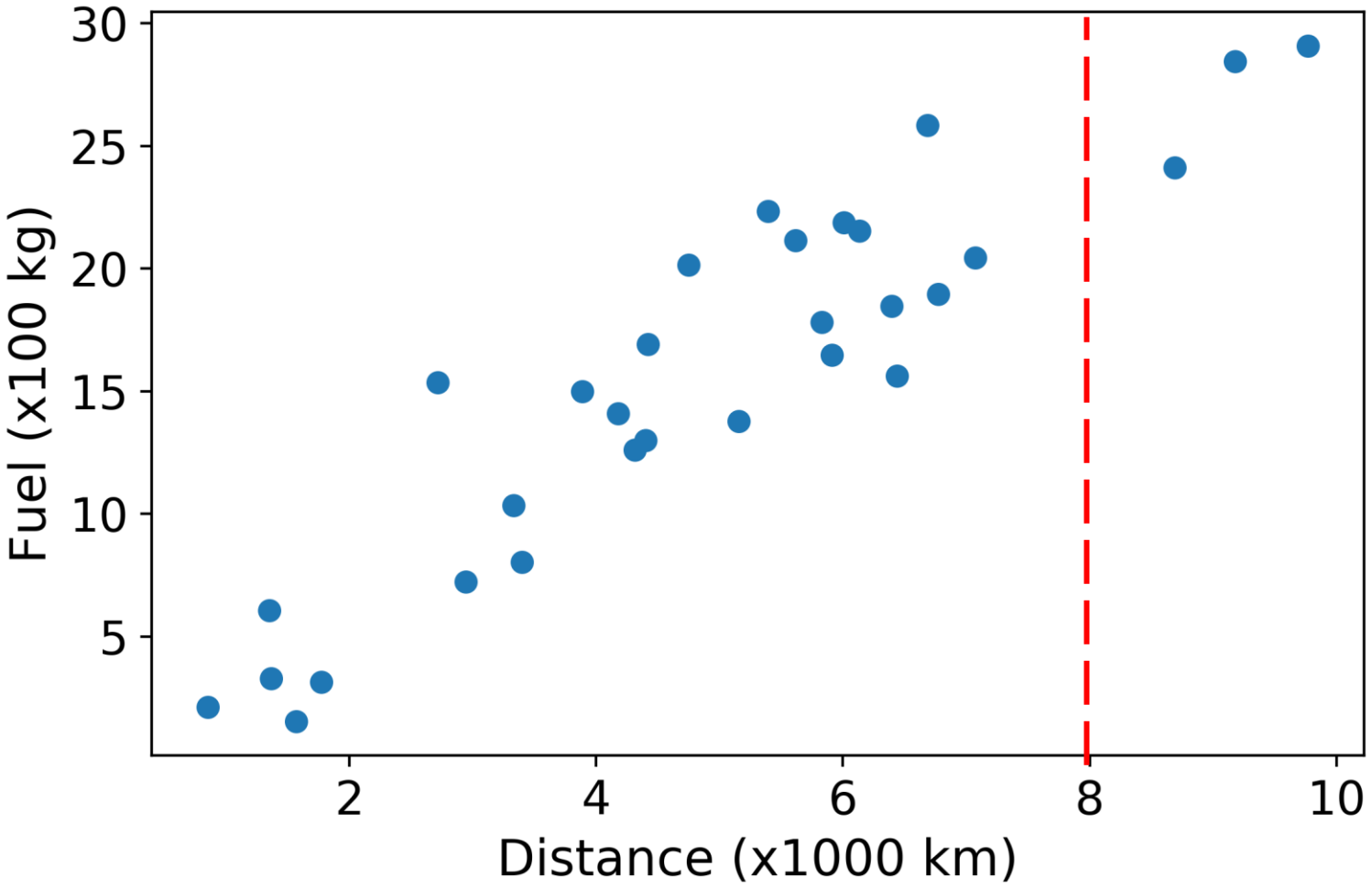


Image source: https://www.vectorstock.com/royalty-free-vector/fuel-station-cartoon-silhouette-vector-15209899

Image source: https://www.open.edu/openlearn/science-maths-technology/engineering-and-technology/aeroplane-design-and-engineering

Past data on distance and fuel



Red line represents the different fuel values for the distance the new flight will travel.

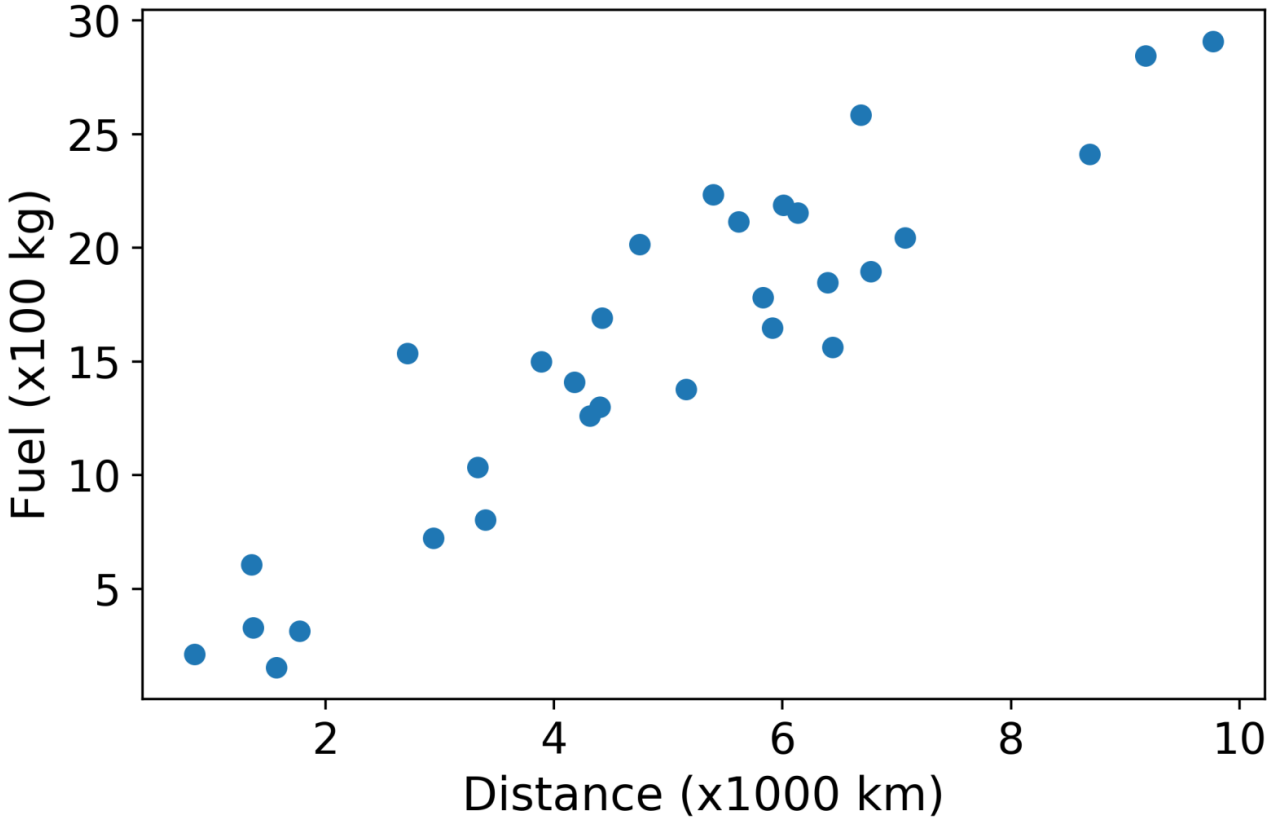
If we can find a function relating distance to fuel, then we are done.

Linear regression

- **Assume** a **linear relationship** between distance (X) and fuel (Y)
  - This is our model assumption

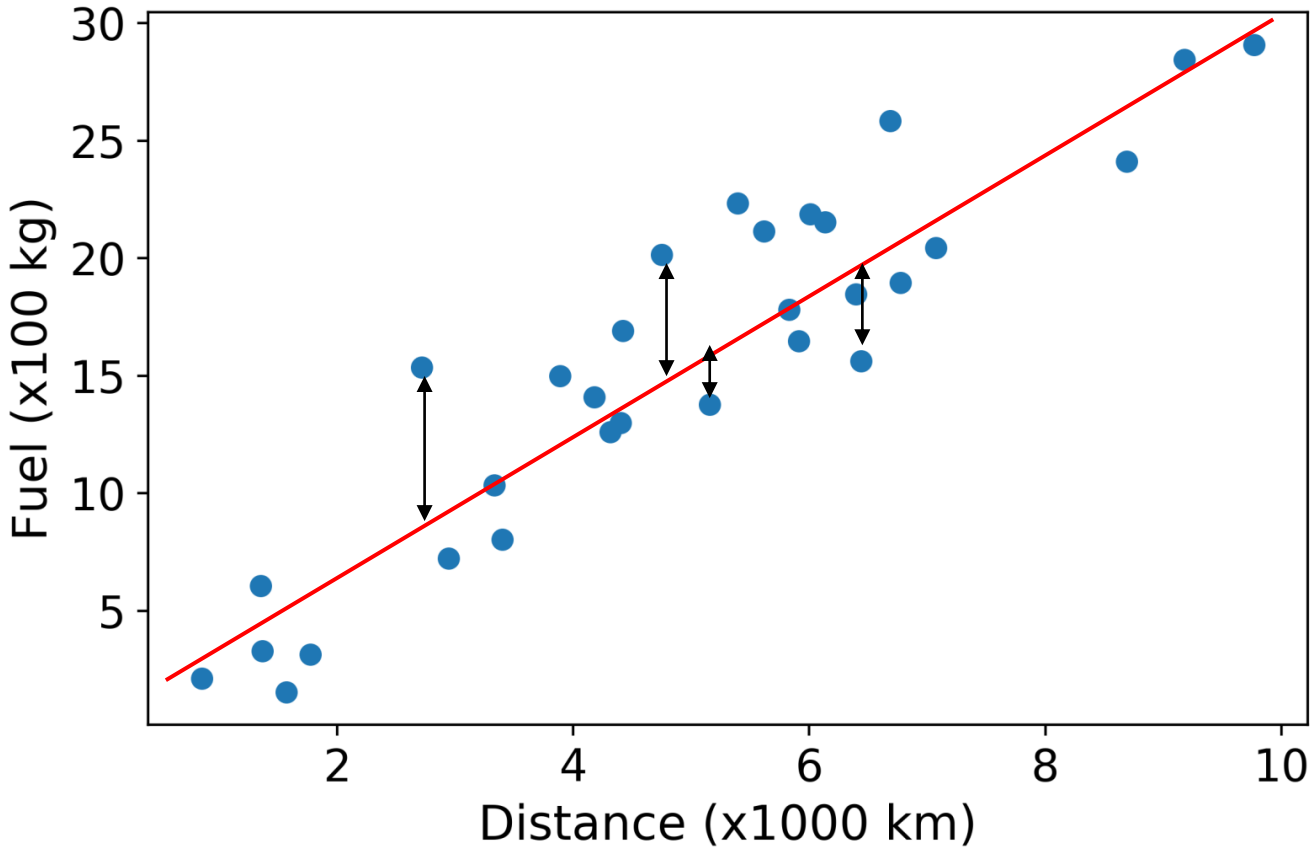
$$Y = \beta X + \alpha$$

- Is the above model sufficient to find relation between D and F?
  - All the points are not collinear, so no solution exists for the above model
- Need something more….



Noise

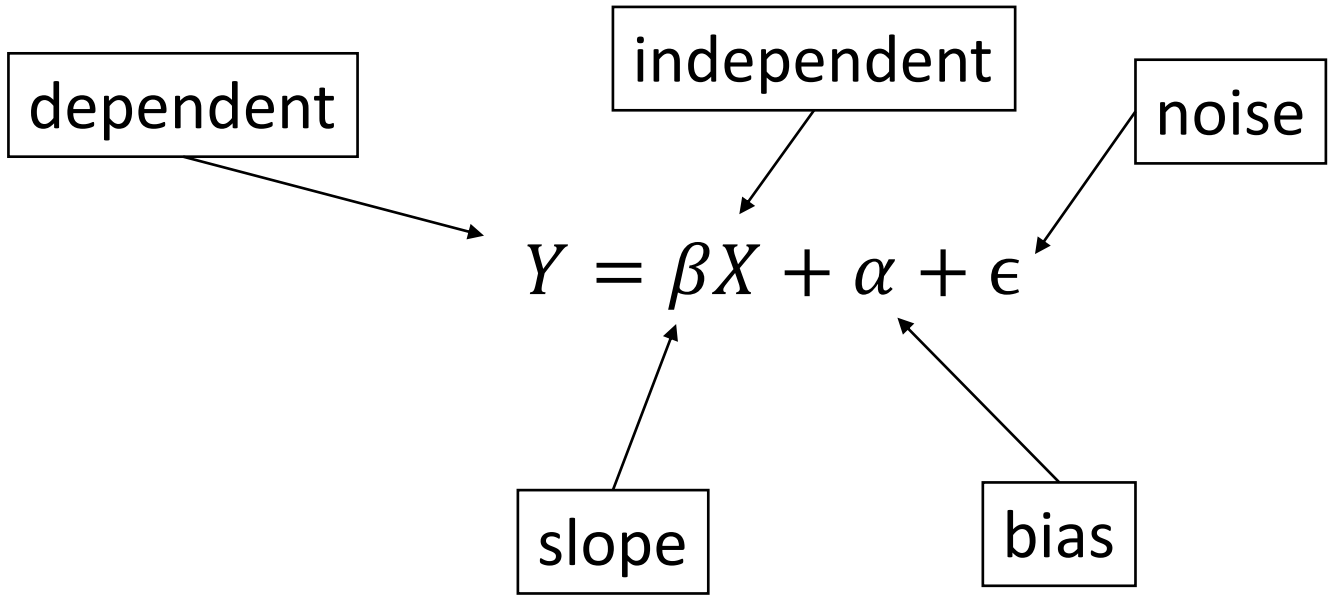
- There is some variation not explained by the model
- Variation is different for each sample
- If we had a random variable to account for this variation, our model could fit
- Above is called noise
- Could represent measurement error as well as variation from understood/unaccounted for variables
  - Weather, number of passengers etc.



Hypothetical linear relation between distance and fuel

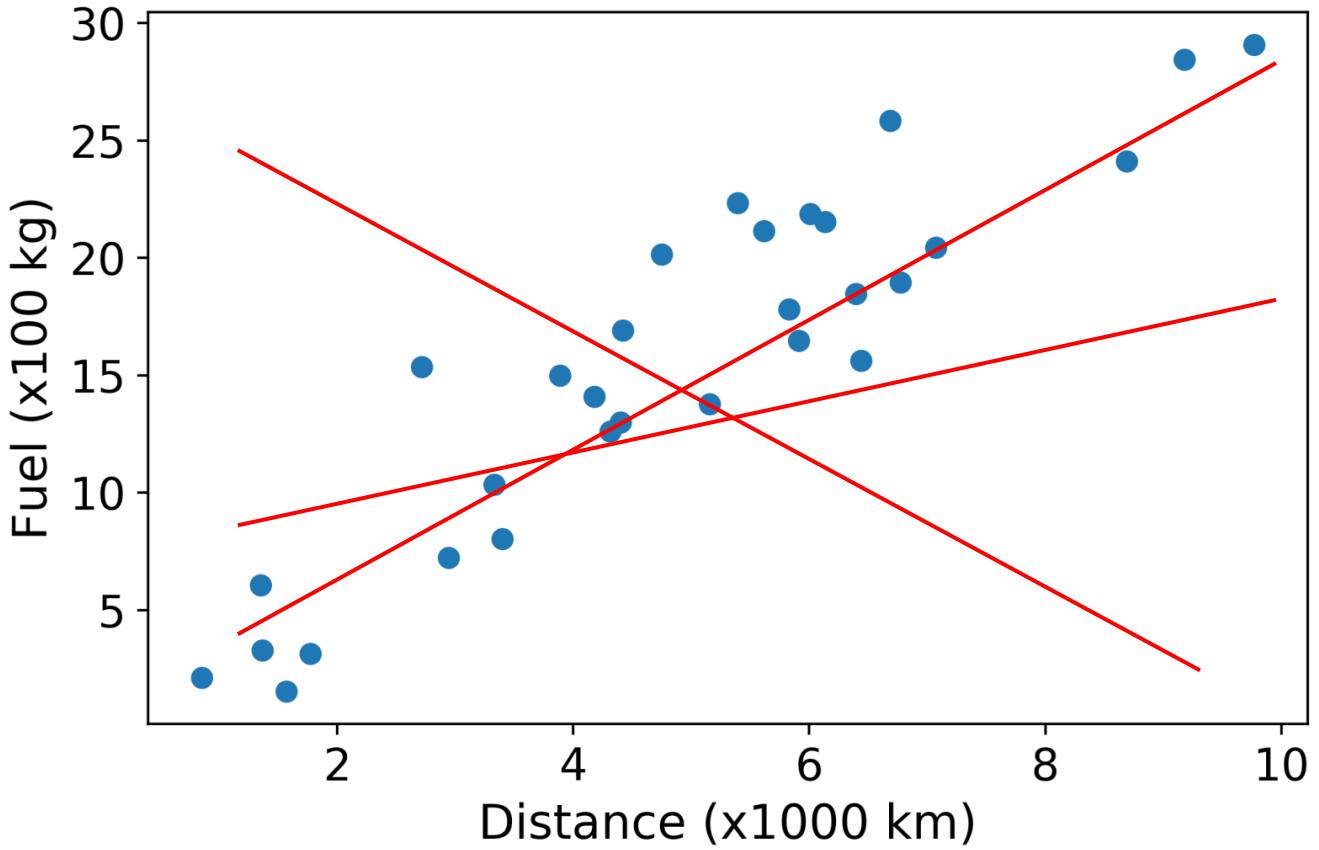
Linear regression

Model with noise



- Noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is Gaussian distributed with zero mean and variance σ^2
- Noise is independent of for each sample
- Variance of noise is the same for all values of X

How to find β and α?



- Several different possibilities for β and α; which is the best one?
- Could be one that maximizes the probability of observation the data

Computing likelihood of the data

$$Y = \beta X + \alpha + \epsilon$$

Say, $\beta = 2$ and $\alpha = 2$, and training data is $(x_i, y_i), i = 1, \dots, N$. What is the probability of data under the above model?

$p(y_1, y_2, \dots, y_n | x_1, \dots, x_n; \beta = 2, \alpha = 2)$ Independence of the samples

$= p(y_1 | x_1; \beta = 2, \alpha = 2) p(y_2 | x_2; \beta = 2, \alpha = 2) \dots p(y_n | x_n; \beta = 2, \alpha = 2)$

$$= \prod_{i=1}^{n} p(y_i | x_i; \beta = 2, \alpha = 2)$$

Computing likelihood of the data

What is $p(y_i|x_i; \beta = 2, \alpha = 2)$?

Probability density of observing $Y = y_i$ when $X = x_i$ and $\beta = 2, \alpha = 2$.

$$y_i = 2x_i + 2 + \epsilon$$

$$y_i \sim \mathcal{N}(2x_i + 2, \sigma^2)$$

$$p(y_i|x_i; \beta = 2, \alpha = 2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - (2x_i + 2)\right)^2}{2\sigma^2}\right)$$

Computing likelihood of the data

Substituting,

$$p(y_1, y_2, \dots, y_n | x_1, \dots, x_n; \beta = 2, \alpha = 2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - (2x_i + 2)\right)^2}{2\sigma^2}\right)$$

Likelihood

Take log on both sides,

$$\log(p(y_1, y_2, \dots, y_n | x_1, \dots, x_n; \beta = 2, \alpha = 2)) = \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - (2x_i + 2)\right)^2}{2\sigma^2}\right)\right)$$

$$\mathcal{L}(\beta = 2, \alpha = 2) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{\left(y_i - (2x_i + 2)\right)^2}{2\sigma^2}$$

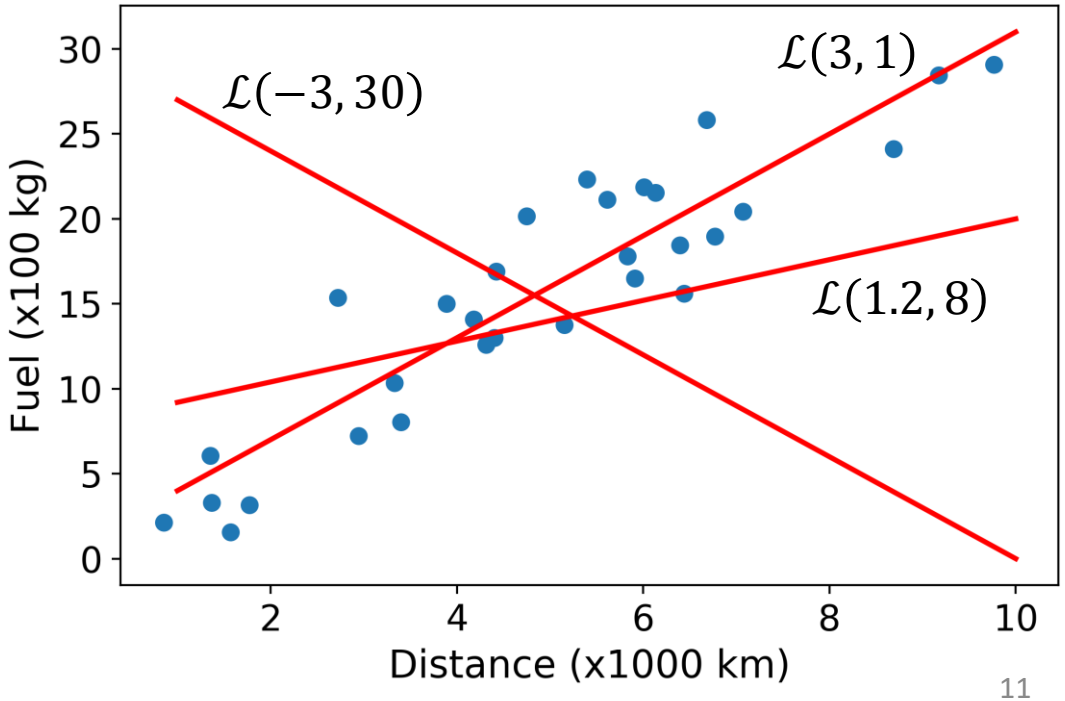
Log-likelihood

Log likelihood of data

Above computation was for $\beta = 2, \alpha = 2$. In general,

$$\mathcal{L}(\beta, \alpha) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{\left(y_i - (\beta x_i + \alpha)\right)^2}{2\sigma^2}$$

- $\mathcal{L}(\beta, \alpha)$ gives the likelihood (probability) of observing the data for different values of β and α
- Picking the best line is the same as choosing the values of β and α that maximizes $\mathcal{L}(\beta, \alpha)$



Calculating β and α

Find β and α that maximize $\mathcal{L}(\beta, \alpha)$

$$\beta^*, \alpha^* = \arg\max_{\beta,\alpha} \mathcal{L}(\beta, \alpha)$$

$$= \arg\max_{\beta,\alpha} \left(-\frac{n}{2}\log(2\pi\sigma^2) \; - \sum_{i=1}^{n} \frac{\left(y_i - (\beta x_i + \alpha)\right)^2}{2\sigma^2} \right)$$

$$\max f(x) = \min -f(x)$$

$$= \arg\min_{\beta,\alpha} \left(\frac{n}{2}\log(2\pi\sigma^2) + \sum_{i=1}^{n} \frac{\left(y_i - (\beta x_i + \alpha)\right)^2}{2\sigma^2} \right)$$

$$= \arg\min_{\beta,\alpha} \sum_{i=1}^{n} \frac{\left(y_i - (\beta x_i + \alpha)\right)^2}{2\sigma^2}$$

$\frac{n}{2}\log(2\pi\sigma^2)$ is a constant

Calculating β and α

Find β and α that maximize $\mathcal{L}(\beta, \alpha)$

$$\beta^*, \alpha^* = \arg\min_{\beta,\alpha} \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - (\beta x_i + \alpha)\right)^2$$

Alternate interpretation: Minimize the squared error between the model prediction ($\hat{y}_i = \beta x_i + \alpha$) and the ground truth ($y_i$)

Formula for β and α

The optimization problem can be solved using standard tools from Calculus.

$$\beta^* = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i) \cdot (\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$\alpha^* = \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

Substituting our training data in the above formulae, we get $\beta = 3.1, \alpha = 0.9$

Fuel to travel 8000km

- Substituting $X = 8$ in the equation, we get

$$\hat{Y} = 3.1 \times 8 + 0.9 = 25.7$$

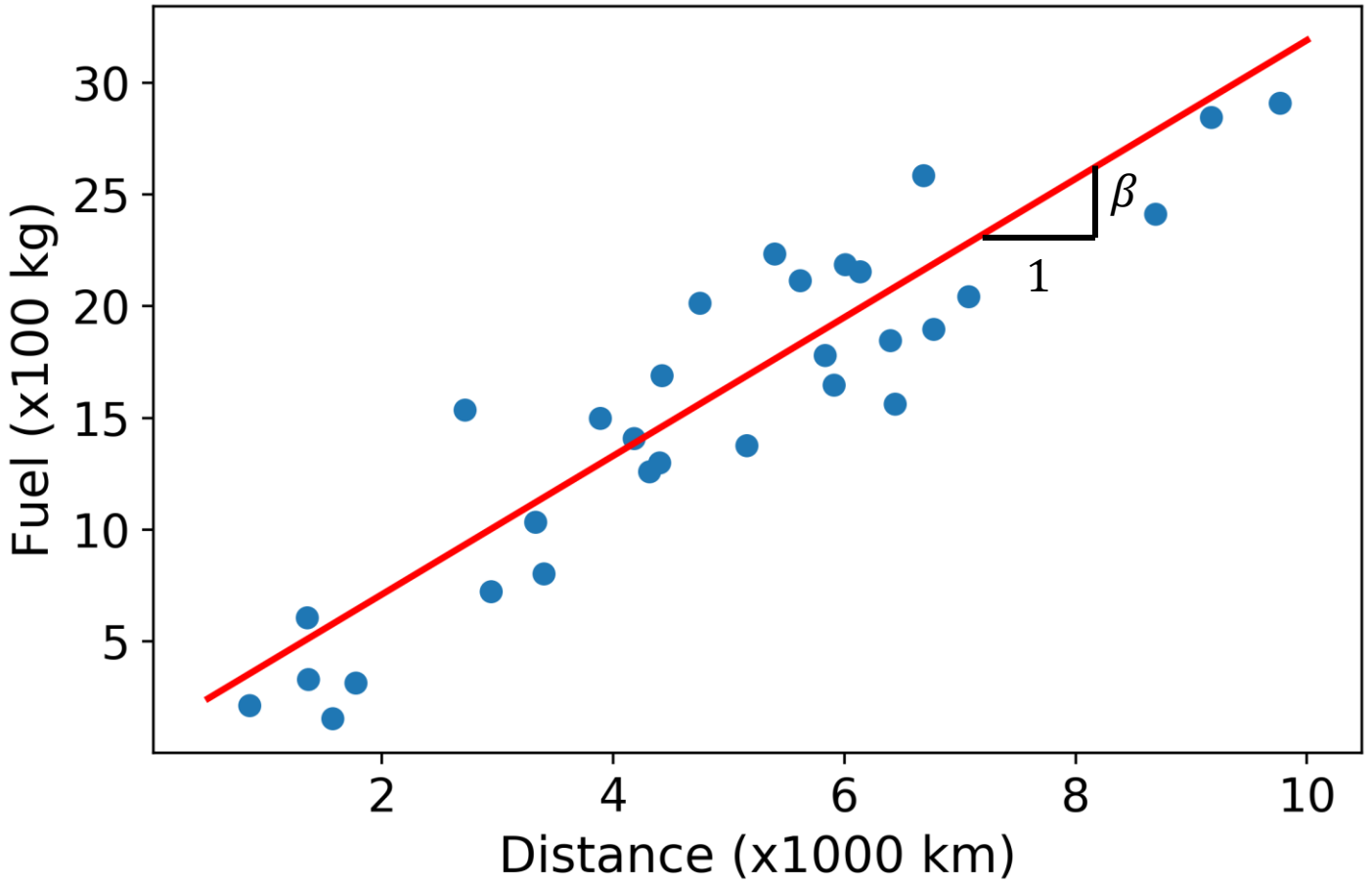
- Fuel needed is 25.7×100kg

Interpretation of the model

$$\hat{Y} = 3.1X + 0.9$$

β: 1000km increase in distance will require 310kg more fuel

α: Amount of fuel required to travel 0km is 90 kg



How well did the model fit: MSE

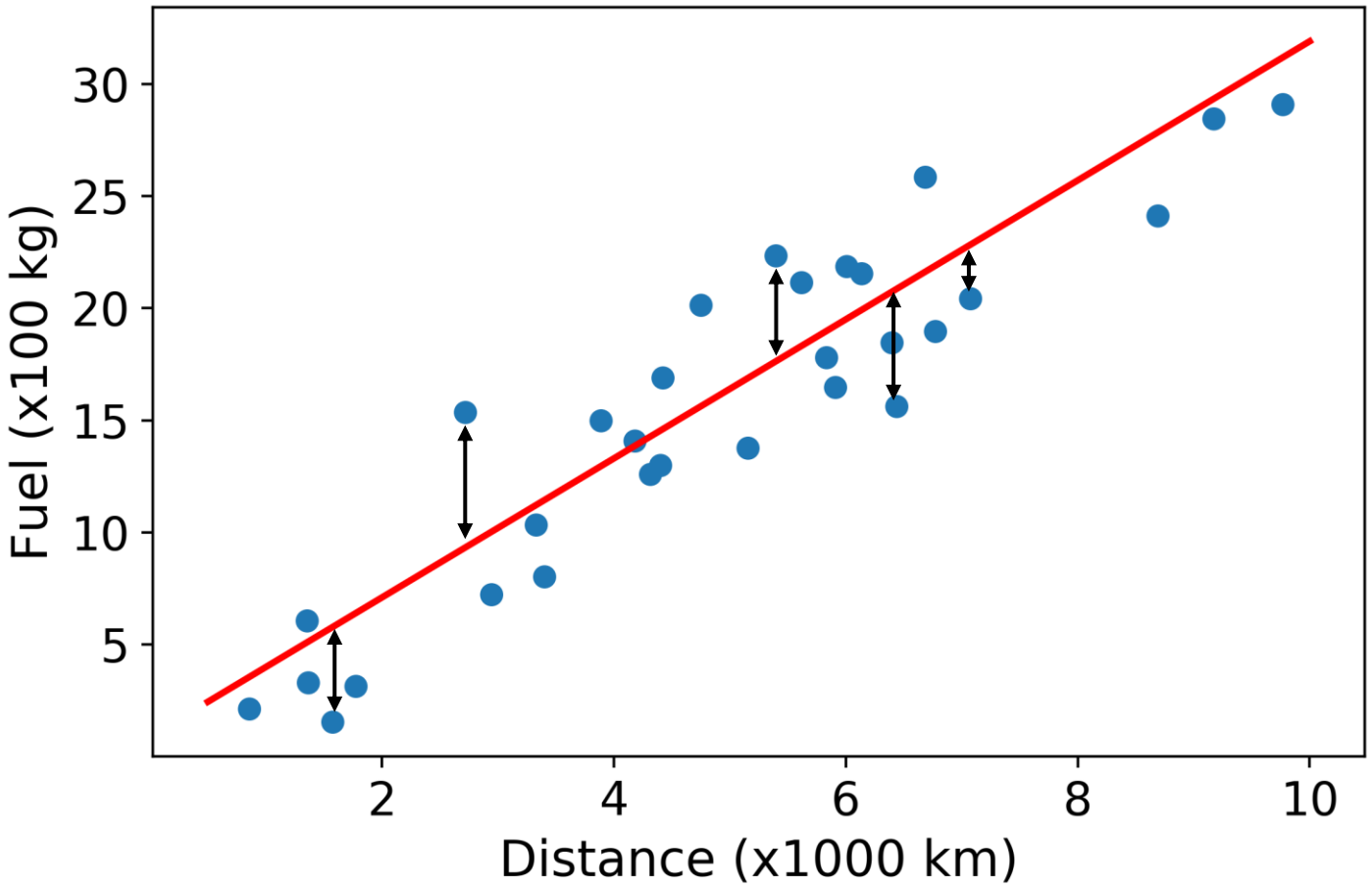
For point x_i

Prediction: $\hat{y}_i$

Ground truth: y_i

MSE: Mean squared error

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



How well did the model fit: R^2

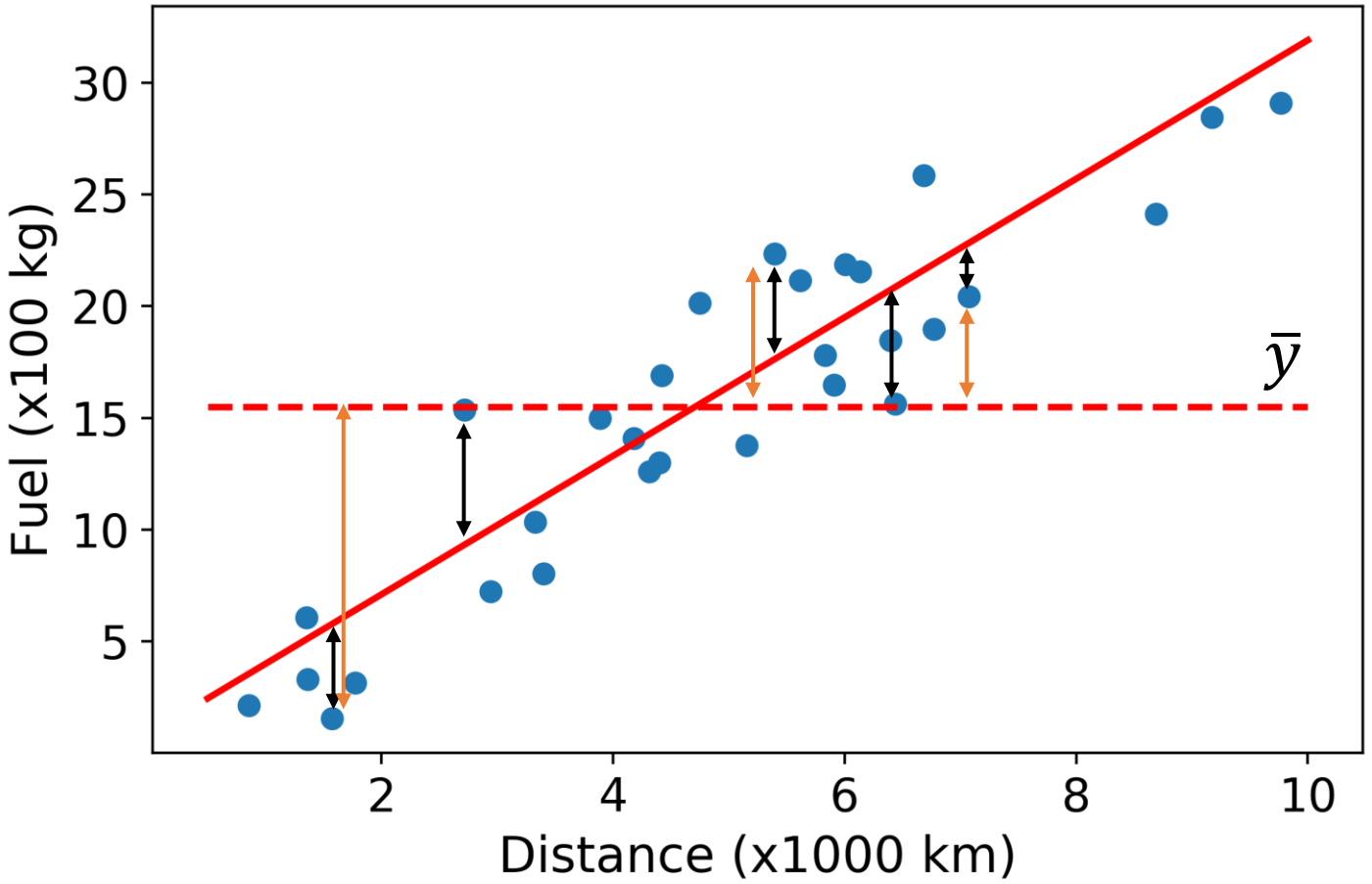
mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Error when prediction is $\bar{y}$ for all values of x (Total Sum of Squares)

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Error when prediction is $\hat{y}_i$ (Sum of Squared Error)

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



$$R^2 = 1 - \frac{SSE}{SST}$$

What is the interpretation of r-squared?
Variance in data explained.

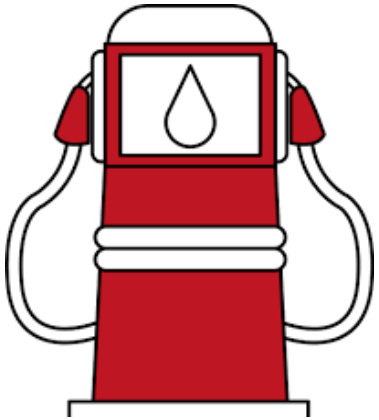
Quick recap

- We started with one independent variable (X; distance), one independent variable (Y; fuel)
- Model: $Y = \beta X + \alpha + \epsilon$
- Fit β and α according to the training data
- Interpret the model
- Evaluate the model fit: $\hat{Y} = \beta X + \alpha$

What it takes to fly….

Previous experience suggests that the amount of fuel depends on multiple factors:

- Distance
- Payload
- Pilot's experience

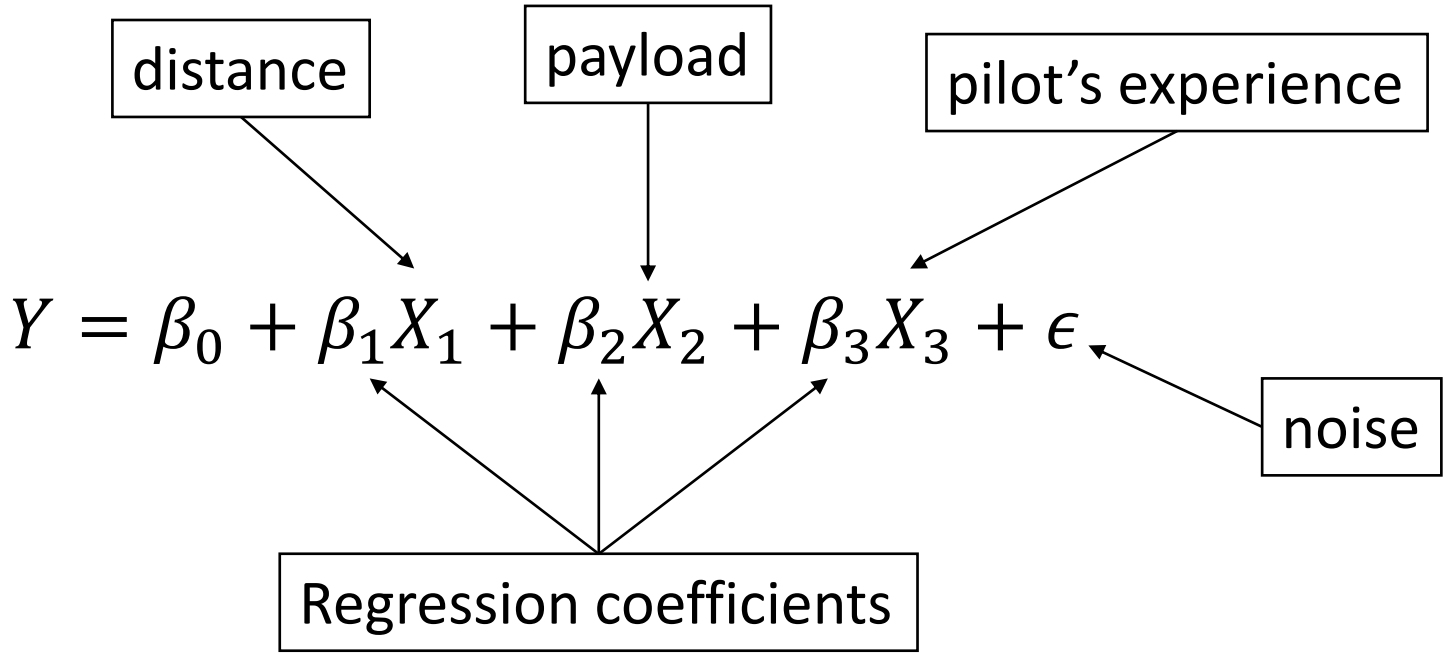


How can this information be used to estimate the fuel requirement in the linear regression model?

Image source: https://www.vectorstock.com/royalty-free-vector/fuel-station-cartoon-silhouette-vector-15209899

Image source: https://www.open.edu/openlearn/science-maths-technology/engineering-and-technology/aeroplane-design-and-engineering

Multivariate linear regression



- Regression coefficients and intercept (bias) are the model parameters
- Model fitting, interpretation, and evaluation is done in the same way as for univariate linear regression
- What information does comparison of regression coefficient provide?

Interpreting regression coefficients

Assume that the features (independent variables) are in the same scale and the regression model after fitting to the training data gives:

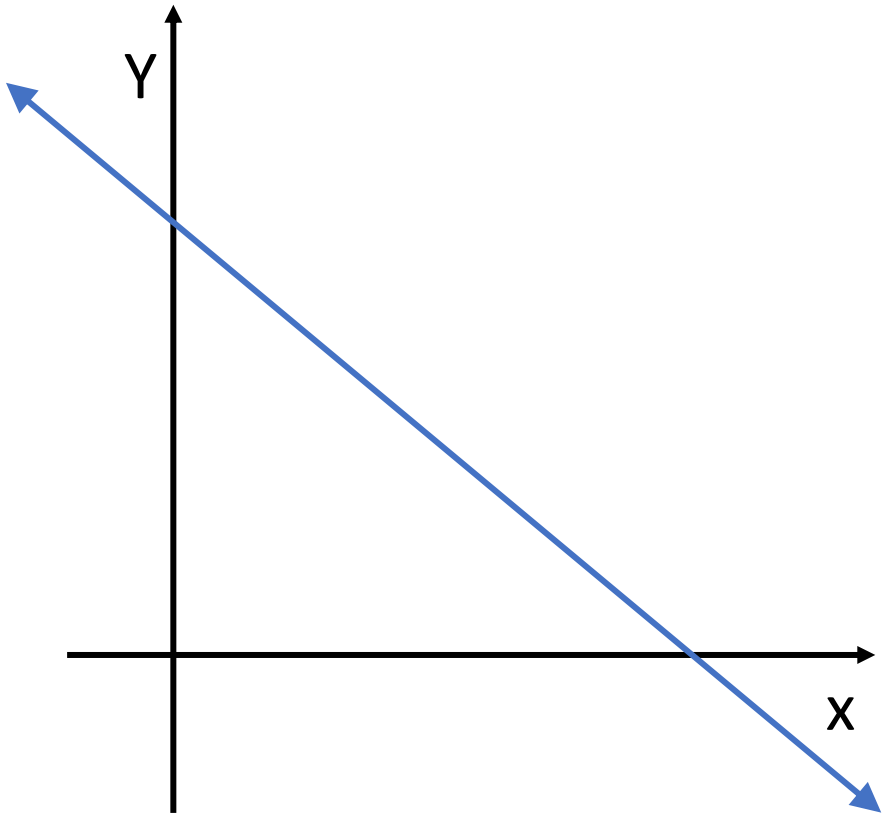
$$Y = 0.9 + 3.1X_1 + 0.8X_2 - 0.2X_3 + \epsilon$$

What do the coefficient values imply about the importance of different features in the fuel requirement?

Logistic Regression

Linear regression like model for classification…

- We got was a nice way of combining variables to get an output with linear regression
  - Gave higher weight to more important features
  - Importance was learned from the data
- Can a similar model wherein a linear combination of the features is taken be used in the classification setting?
- Challenge:
  - The output of linear regression lies between $(-\infty, \infty)$
  - Classification is discrete and binary {0,1}



Ways to tackle the challenges…

Idea for solution to challenges

1. Relate the classification outcome to the result of a coin toss
   - Example: assign label 0 if coin lands Tails and label 1 if coin lands Heads

2. Probability of the heads in the coin toss relates to the independent variables by a linear model
   - We just need a way of converting the $(-\infty, \infty)$ values to lie between $(0,1)$

Logit function

Consider,

$$\mathrm{f(p)} = \log\left(\frac{p}{1-p}\right)$$

Logit function

Logit function takes a value in (0,1) as input and outputs a value in $(-\infty, \infty)$

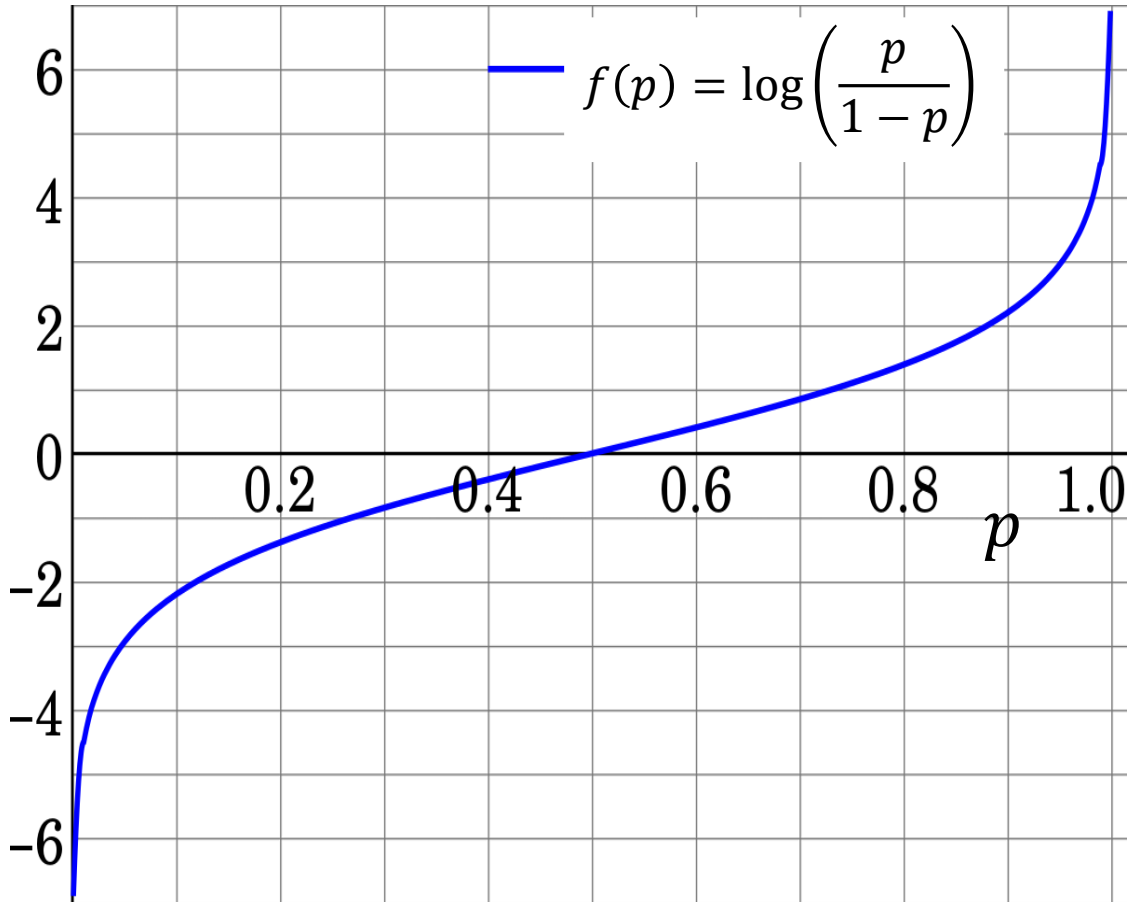


Image source: https://en.wikipedia.org/wiki/Logit

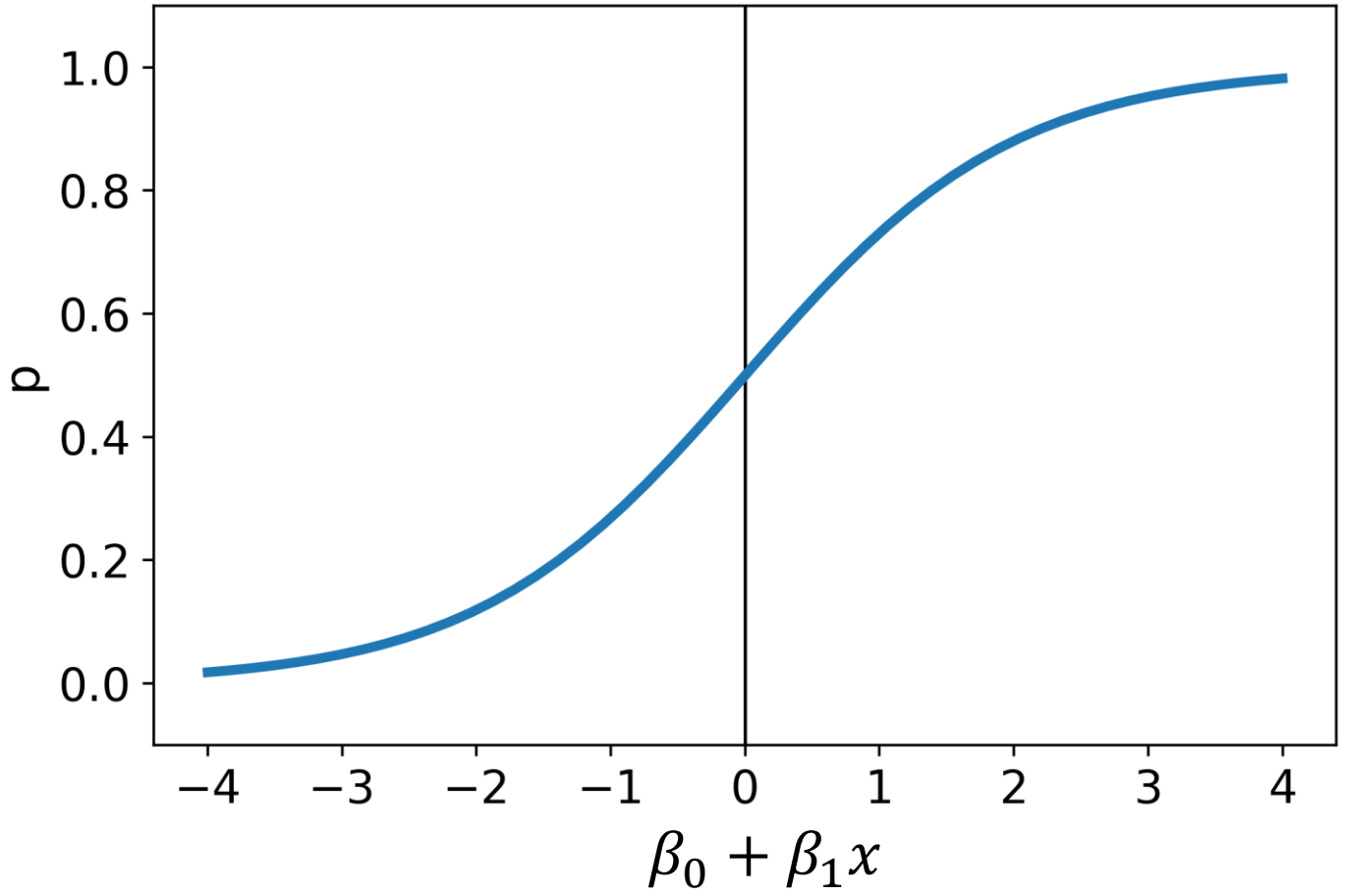
Logistic regression

Let $p = P(Y = 1|x)$

Assume logit function is a linear function of independent variable (feature)

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

$$p = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}$$



Multiple variables and training

If there are multiple features, the above equation can be easily extended. For example, with 3 features,

$$p = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3)}$$

- Training the logistic regression classifier means finding the parameters $\beta_0, \beta_1, \ldots$ given training data (X_i, Y_i) where $X_i = [x_{i,1}, x_{i,2} \ldots], i = 1, \ldots, n$
- Standard optimization tools can be used to train the model

Logistic regression example

Recall Ms. Orange Seller's task of finding whether an orange is navel (label=1) or clementine (label=0). She trains a logistic regression model with size (x_1) and weight (x_2) as features. The trained model has regression coefficient $\beta_1 = 0.1, \beta_2 = 0.2, \beta_0 = -5$.

What is $P(Y = 1|X)$?

$$P(Y = 1|X) = \frac{1}{1 + \exp(-(-5 + 0.1x_1 + 0.2x_2))}$$

Logistic regression example

Recall Ms. Orange Seller's task of finding whether an orange is navel (label=1) or clementine (label=0). She trains a logistic regression model with size (x_1) and weight (x_2) as features. The trained model has regression coefficient $\beta_1 = 0.1, \beta_2 = 0.2, \beta_0 = -5$.

An orange has size=5 and weight=20 i.e., $X = [5, 20]$. What type of orange is it?

$$P(Y = 1|X) = \frac{1}{1 + \exp\big(-(-5 + 0.1 \times 5 + 0.2 \times 20)\big)} = 0.38$$

$$P(Y = 0|X) = 1 - P(Y = 1|X) = \ 0.62$$

$P(Y = 0|X) > P(Y = 1|X)$; so orange is label=0 i.e., clementine

Logistic regression example

Recall Ms. Orange Seller's task of finding whether an orange is navel (label=1) or clementine (label=0). She trains a logistic regression model with size (x_1) and weight (x_2) as features. The trained model has regression coefficient $\beta_1 = 0.1, \beta_2 = 0.2, \beta_0 = -5$.

An orange has size=10 and weight=40 i.e., $X = [10, 40]$. What type of orange is it?

$$P(Y = 1|X) = \frac{1}{1 + \exp\left(-(-5 + 0.1 \times 10 + 0.2 \times 40)\right)} = 0.98$$

$P(Y = 1|X) > 0.5$; so orange is label=1 i.e., navel

Questions?