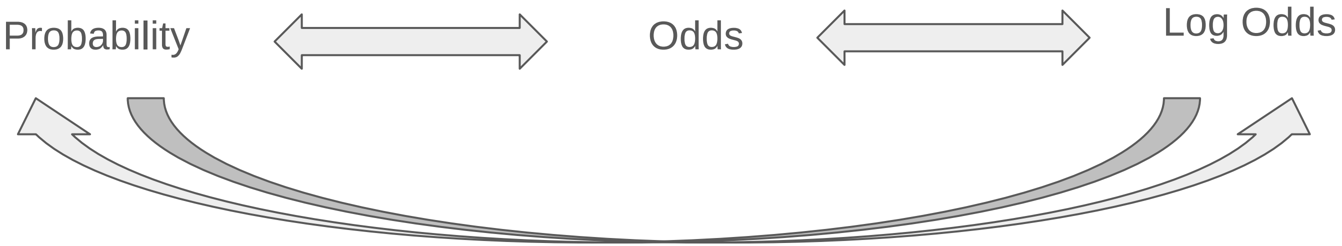
Logistics Regression

Learning Objectives

- Describe probability, odds and log odds
- Discuss logit and sigmoid functions
- Explain maximum likelihood estimation
- Explain beta coefficient with demonstration
- Demonstrate logistic regression model in python using statsmodel

Probability, Odds, Log Odds

Overview



Example:

One want to go jogging when there is **no rain**.

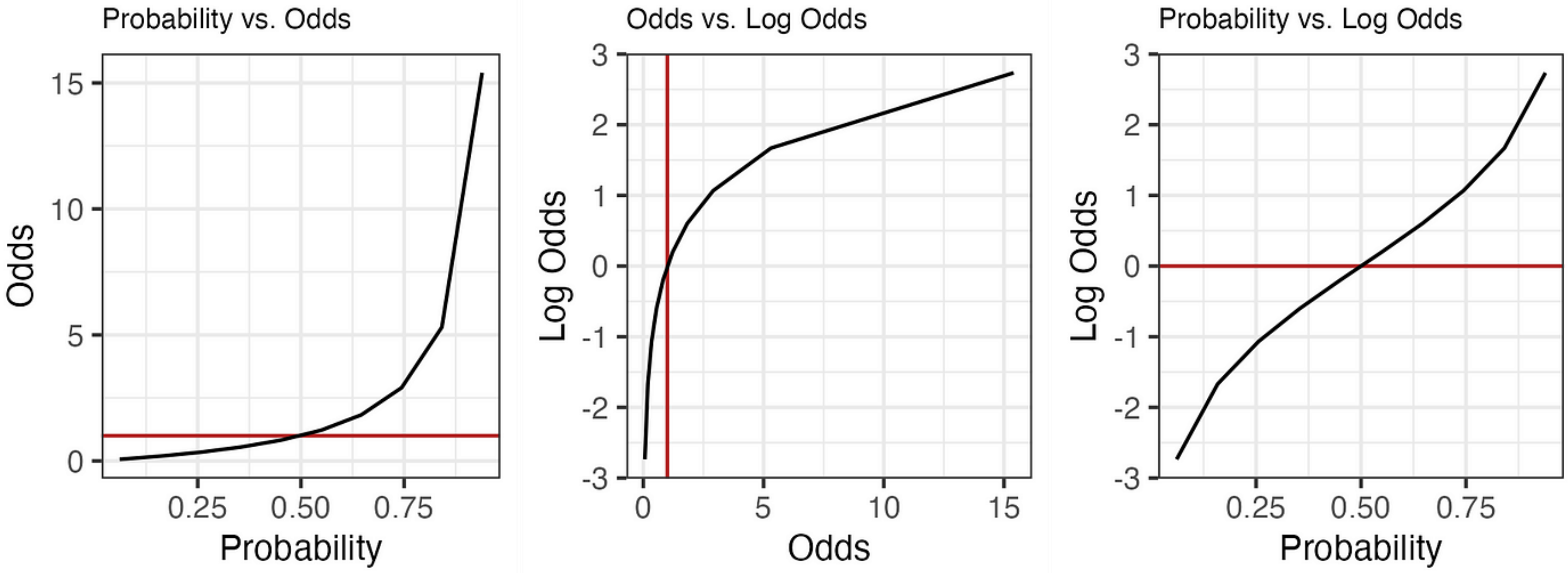
According to the weather forecast, the **probability** for rain is 30%.

The **odds** for no-rain are 0.7 / 0.3 = 2.3 = 2.3 : 1.

The **log odds** are the logarithm of the odds = log(2.3) = 0.36.

Probability, Odds, Log Odds

$$odds = \frac{p}{(1-p)} \qquad \log(odds) = log\left(\frac{p}{1-p}\right)$$



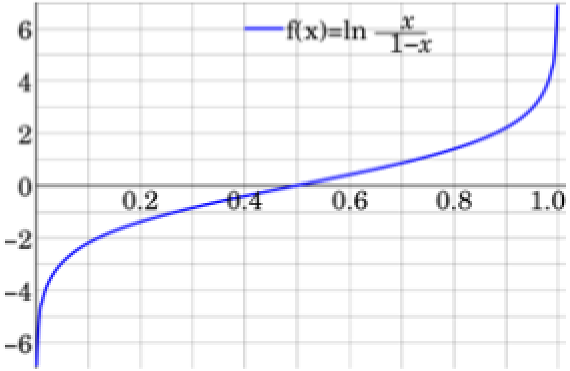
Logit and Sigmoid

Logit Function

Get real numbers from probability.

$$odds = \frac{p}{1-p}$$

$$log(odds) = log(\frac{p}{1-p}) = logit\ function$$

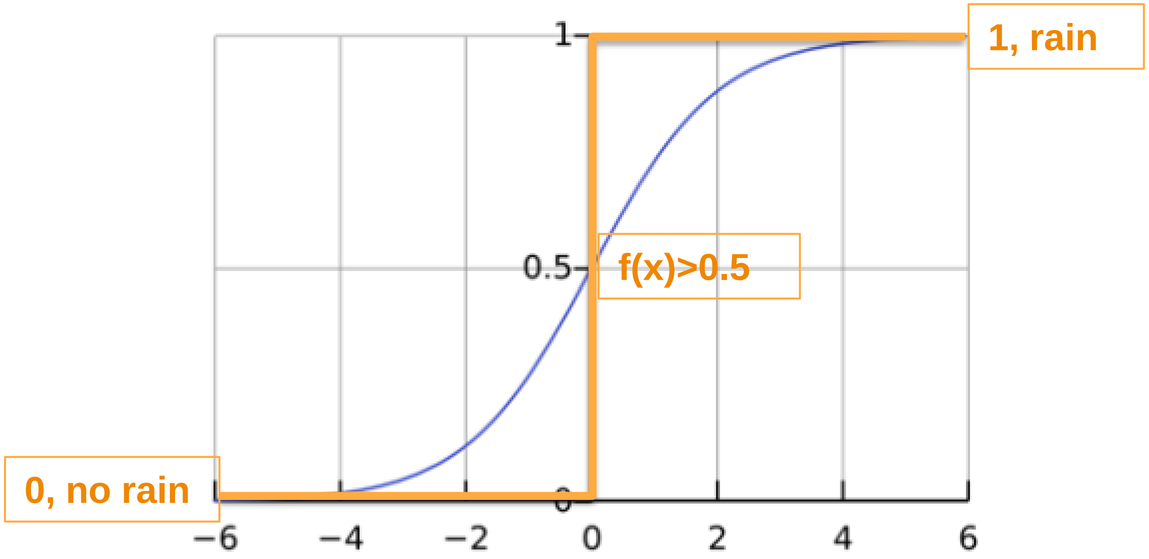


Sigmoid Function

Get probability from real numbers.

$$f(x) = \frac{1}{1 + e^{-x}}$$

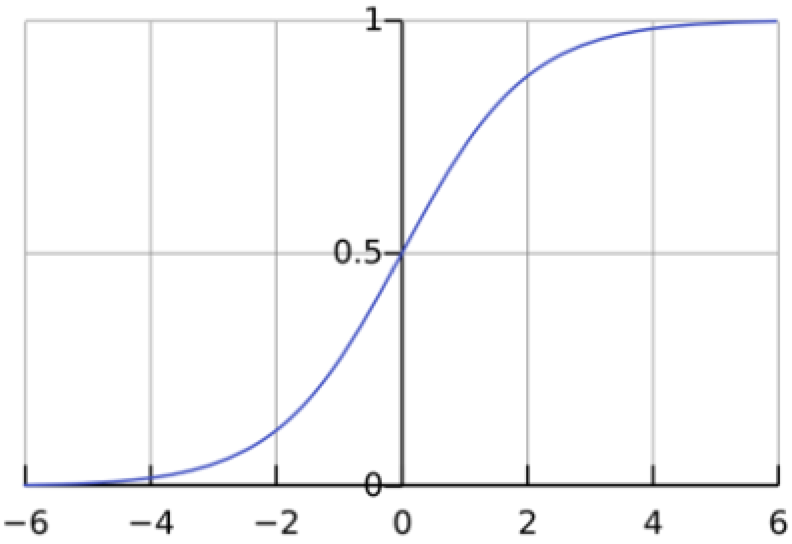
e… Euler's number (exp)



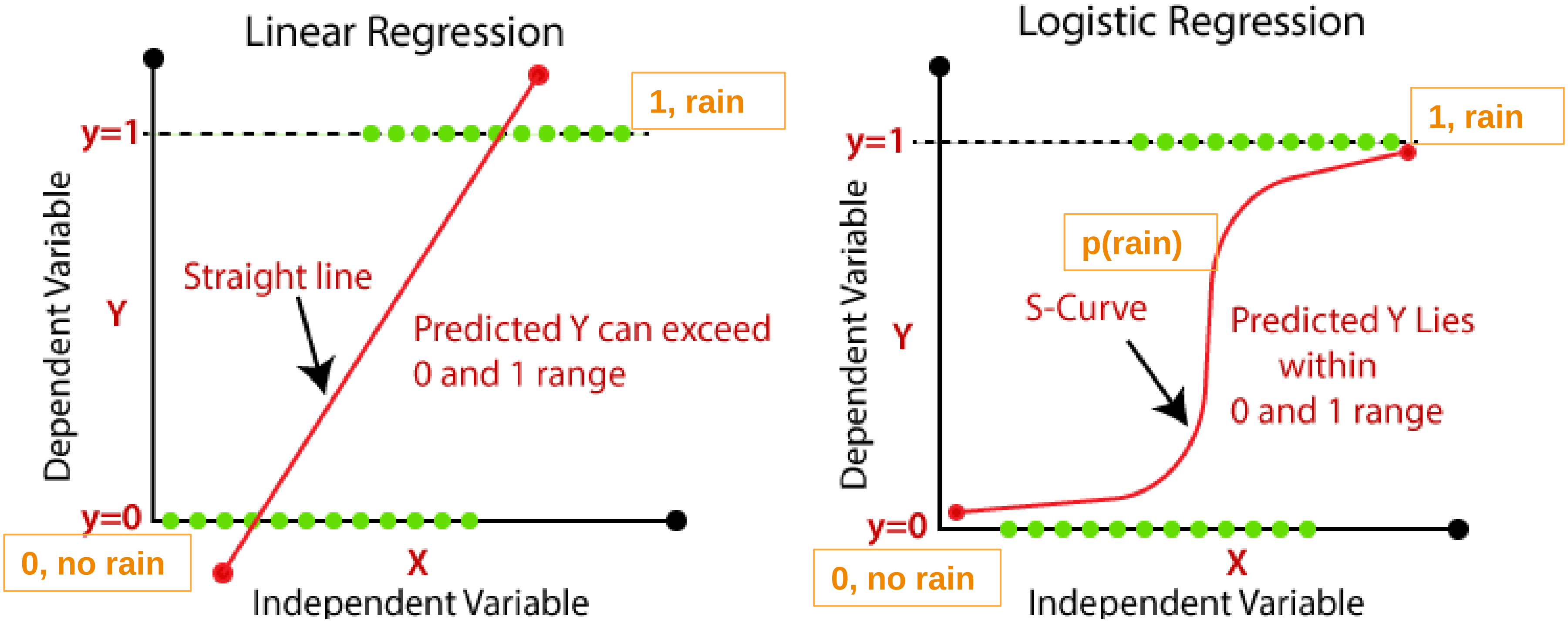
Logistic Regression

Plug-in linear regression equation!

$$p(x) = \frac{1}{1 + e^{-\underbrace{(\beta_0 + \beta_1 x)}_{R}}}$$



Logistic Regression



https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning

Maximum Likelihood Estimation

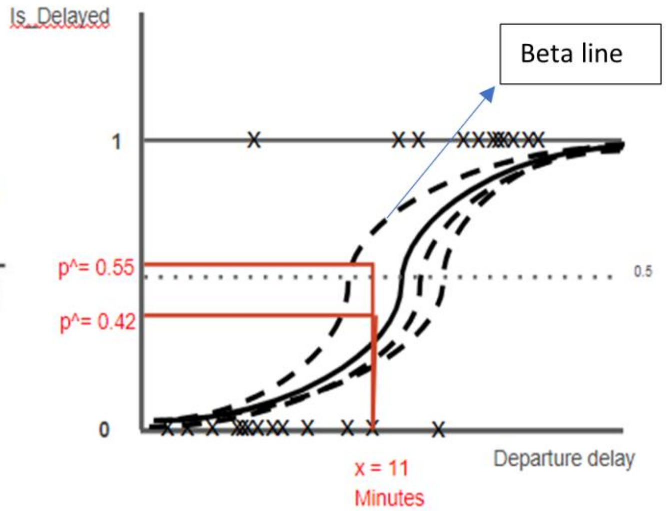
Probability of Variables

- Sigmoid Function

$$f(x) = \frac{1}{1+e^{-x}}$$

- Logistic Regression Model

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

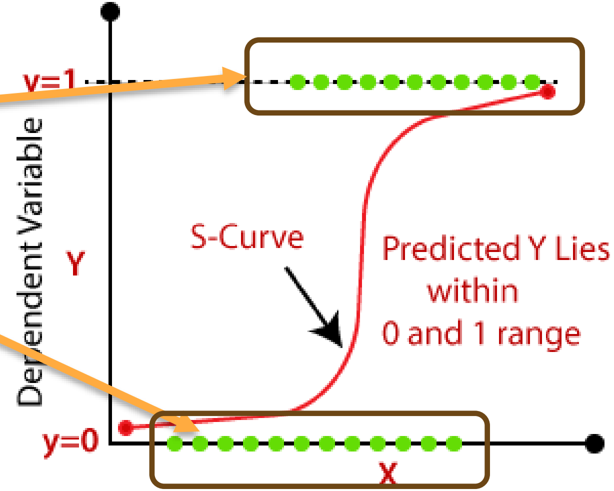


Loss Function

Try to estimate beta such as the product of all probabilities for classes labeled as “1” is largest and for classes labeled as “0” is smallest.

$$for\ samples\ labelled\ as\ 1: \prod_{s\ in\ y_i=1} p(x_i)$$

$$for\ samples\ labelled\ as\ 0: \prod_{s\ in\ y_i=0} (1-p(x_i))$$



Maximize, p(x)

Minimize, p(x) or Maximize, 1-p(x)

Likelihood Function

Goal: Find Beta to **maximize** this function.

$$L(\beta) = \prod_{s\ in\ y_i = 1} p(x_i) \ * \prod_{s\ in\ y_i = 0} (1 - p(x_i))$$

Likelihood Function

Goal: Find beta so that this function gets maximized.

$$l(\beta) = \sum_{i=1}^{n} y_i \beta x_i - log\left(1 + e^{\beta x_i}\right)$$

Beta Coefficient

Coefficient of Linear Regression

The coefficient β associated with a variable X is the **expected change in log odds** of having the outcome Y per unit change in X.

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

x1	x2	xp	
20	2	...	0.234
-14	1	...	0.987
191	2	…	0.456
…	…	…	…

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

Case 1: Input Variable is Numeric

An increase of 1 minute in departure delay multiplies the odds of arrival_delay by 1.19.

An increase of 1 minute in departure delay is associated with an increase of 19% in the odds of arrival_delay.

departure_delay_minutes	arrival_delay_15
12	0.234
35	0.987
16	0.456
…	…

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   30
Model:                          Logit   Df Residuals:                       28
Method:                           MLE   Df Model:                            1
Date:                Mon, 17 Oct 2022   Pseudo R-squ.:                  0.3818
Time:                        11:10:32   Log-Likelihood:                -11.328
converged:                       True   LL-Null:                       -18.326
Covariance Type:            nonrobust   LLR p-value:                 0.0001833
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.4786      0.798     -3.107      0.002      -4.042      -0.915
x1             0.1728      0.060      2.895      0.004       0.056       0.290
==============================================================================
```

e^β = e^0.1728 = 1.19

$$odds = \frac{p}{(1-p)}$$

Case 2: Input Variable is Numeric

Changing from one ordinal level to the next multiplies the odds of arrival delay by 1.19.

Going 1 level up of departure delay s associated with an increase of 19% in the odds of arrival delay.

departure_delay_bin	arrival_delay_15
1	0.234
3	0.987
2	0.456
…	…

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   30
Model:                          Logit   Df Residuals:                       28
Method:                           MLE   Df Model:                            1
Date:                Mon, 17 Oct 2022   Pseudo R-squ.:                  0.3818
Time:                        11:10:32   Log-Likelihood:                -11.328
converged:                       True   LL-Null:                       -18.326
Covariance Type:            nonrobust   LLR p-value:                 0.0001833
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.4786      0.798     -3.107      0.002      -4.042      -0.915
x1             0.1728      0.060      2.895      0.004       0.056       0.290
==============================================================================
```

$$e^\beta = e^{0.1728} = 1.19$$

$$odds = \frac{p}{(1-p)}$$

One Hot Encoder

Same interpretation as binary for each column.

departure_delay_reason	arrival_delay_15
2	0.234
1	0.987
3	0.456
…	…

reason_1	reason_2	arrival_delay_15
0	1	0.234
1	0	0.987
0	0	0.456
…		…

Standard Error

97.5% Confidence Interval for coefficients: β ± 1,95996 × SE = 0.1728 ± 1,95996 × 0.06 = [**0.056, 0.29**].

97.5% Confidence Interval for odds = e^(β ± 1,95996 × SE) = e^(0.1728 ± 1,95996 × 0.06) = **[1.06, 1.34]**.

departure_delay_15	arrival_delay_15
0	0.234
1	0.987
1	0.456
…	…

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   30
Model:                          Logit   Df Residuals:                       28
Method:                           MLE   Df Model:                            1
Date:                Mon, 17 Oct 2022   Pseudo R-squ.:                  0.3818
Time:                        11:10:32   Log-Likelihood:                -11.328
converged:                       True   LL-Null:                       -18.326
Covariance Type:            nonrobust   LLR p-value:                 0.0001833
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.4786      0.798     -3.107      0.002      -4.042      -0.915
x1             0.1728      0.060      2.895      0.004       0.056       0.290
==============================================================================
```

Significance

Same as linear regression (decision threshold to reject the null hypothesis that the coefficient has no effect on Y).

Don't judge on p alone - take a thorough look at the data and conduct an exploratory data analysis!

departure_delay_15	arrival_delay_15
0	0.234
1	0.987
1	0.456
…	…

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   30
Model:                          Logit   Df Residuals:                       28
Method:                           MLE   Df Model:                            1
Date:                Mon, 17 Oct 2022   Pseudo R-squ.:                  0.3818
Time:                        11:10:32   Log-Likelihood:                -11.328
converged:                       True   LL-Null:                       -18.326
Covariance Type:            nonrobust   LLR p-value:                 0.0001833
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.4786      0.798     -3.107      0.002      -4.042      -0.915
x1             0.1728      0.060      2.895      0.004       0.056       0.290
==============================================================================
```

Thank you