Lecture 5 - 10/09/24

Logistic Regression

Classification Problems:

- So far regression problems were → Given X_0 output y_0
- Many times we wanna know which class a data point belong to...

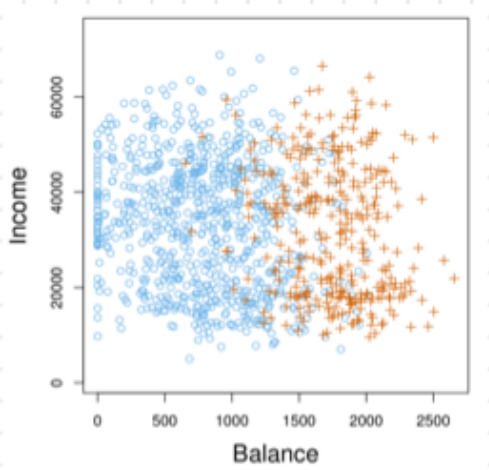
  i.e

  Based on:
  - Annual Income
  - Monthly credit card balance

  } predict whether the individual will go broke or not!

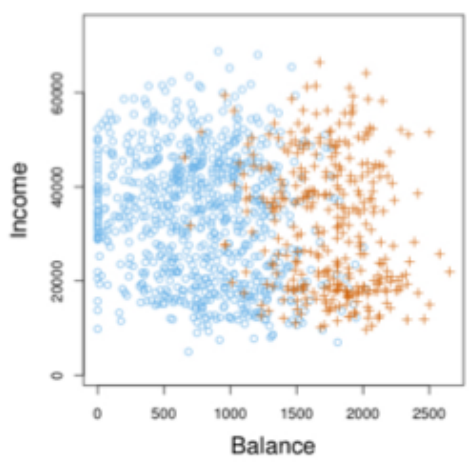
  The outcome of this is binary, either yes or no... this is exactly what logistic regression predicts.

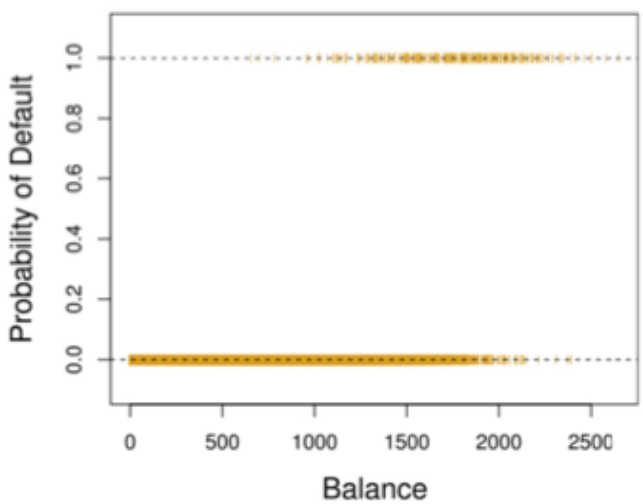
- Linear regression predicts continuous values
- Logistic Regression predicts discrete values (binary values) !!!
- We then want our model to classify data points
  → (classification problem)



- We want a model that guesses the color of the datapoints
- We immediately see we are tryin to classify!
- This is a binary classification

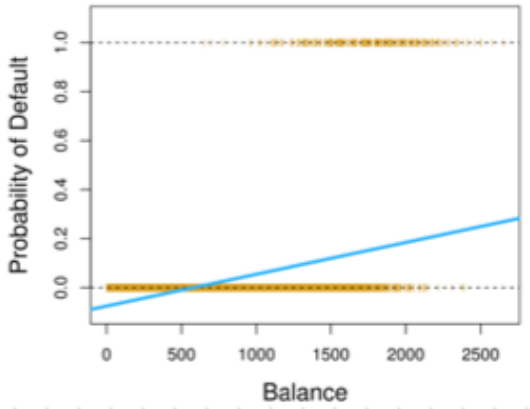
So a popular classification model is **Logistic Regression**





- Using Linear Regression in this example is bad!
- Modeling Linear Regression could give us:

→ $P(X) < 0$ || $P(X) > 1$

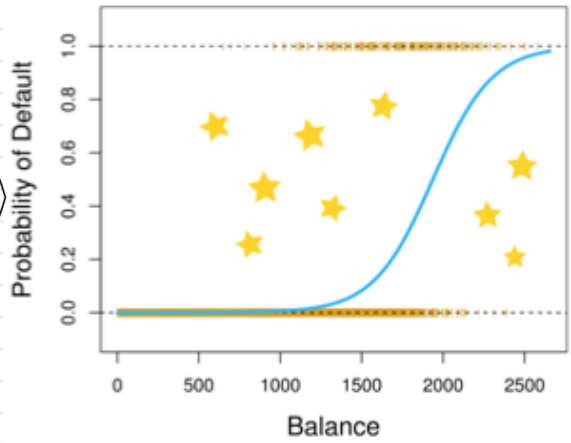


- We are instead seeking a model for $P(X)$:

such that:

Asymptotically gives $P(X) \to 1$

Asymptotically gives $P(X) \to 0$



Logistic Function

- What we are looking for is the **Logistic function:**

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

↳ Sometimes we refer to it as the **Sigmoid function**

↳ The exponent looks like our Linear Regression...

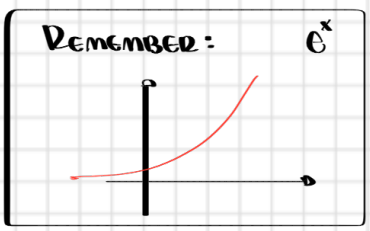
↳ But there are nice properties to it...

What if $\beta_0+\beta_1 X \to \infty$? $\frac{\infty}{\infty} = 1$

What if $\beta_0+\beta_1 X \to -\infty$? $\frac{0}{1+0} = 0$

What if $\beta_0+\beta_1 X \to 0$? $\frac{1}{1+1} = 1/2$

Remember: e^x



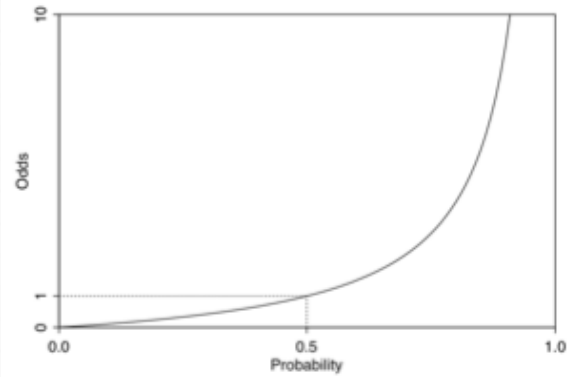
↳ Given this we infer that the logistic function always produces S-shape!!

↳ Sometimes likelihood is thought as:

$\frac{P(X)}{1-P(X)}$ ← Prob of X happening / Prob of X not happening

Probability of an odd

- Odds are positive real numbers
- odds(A) > odds(B) exactly when P(A) > P(B)



Odds of 1 is a bit special in the sense that it has the same prob of that event happening as not happening.

It is a 50-50 chance.

- The odds in two groups can be compared by their odds ratio
- Often use ratio to compare chance of event happening under 2 ≠ conds
- odds = 5 → odds of having cancer is 5 times higher for a smoker than a non

cond 1: patient smokes

cond 2: patient doesn't smoke

- For the logistic function:

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

- Taking the log gives us the log odds / logit:

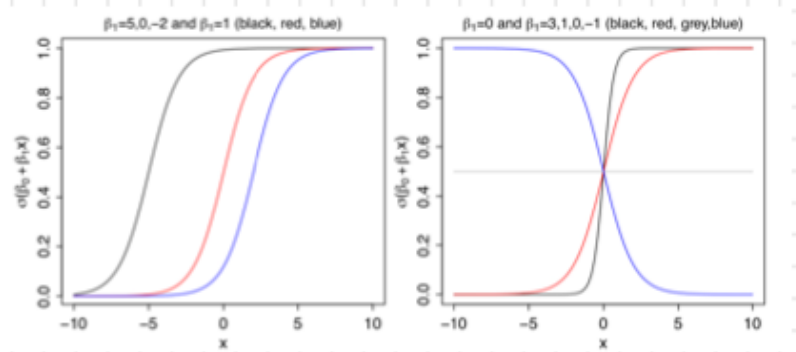
$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0+\beta_1 x$$

Logistic Regression has a logit that is linear in X.

Probabilities as function of feature X

Recall:

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



- We see that β_0/β_1 determine where $P(x) = 0.5$
- Interpretation of coefficients:
  - A unit change in feature X_i means:
    - Change β_i in log-odds
    - Multiplicative change in odds by a factor e^{β_i} where:
      - e^{β_i} is an odds ratio.
      - β_i is a log odds ratio.
  - Given the nonlinear translation between odds & probability, it is hard to communicate the actual change in $P(x)$

BUT:

! LOGIT-TRANSFORMATION IS MONOTONE!

! POSITIVE β_1 GIVES POSITIVE CHANGE IN P(X)!

! A HIGHER β_1 GIVES A STEEPER CURVE!

LOGISTIC REGRESSION WITH MULTIPLE FEATURES

- GIVEN, $x_1, \dots x_n$, LOGISTIC REG. MODELS CONDITIONAL PROB $y=1$ AS:

$$P(y_i = 1 \mid X_i = x) = \frac{e^{x\beta}}{1 + e^{x\beta}}$$

WHICH THEN LEADS TO:

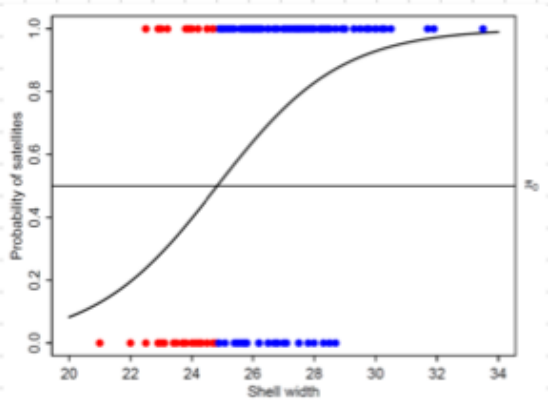
$$\frac{P(x)}{1 - P(x)} = e^{x\beta} \xrightarrow{\text{LOG-ODDS}} \log\left(\frac{P(x)}{1 - P(x)}\right) = x\beta$$

PREDICTION

- TO MAKE A PREDICTION $\hat{y}$:

WE CLASSIFY TO THE CLASS WITH HIGHEST PROBABILITY

$$\hat{y}_i = \begin{cases} 1, & P(y_i = 1 \mid x_i) > 0.5 \\ 0, & P(y_i = 0 \mid x_i) < 0.5 \end{cases}$$



- WE SEE THAT OUR PLOTTED FUNCTION CAN OUTPUT PROBS BETWEEN 0 & 1
- WE DRAW OUR DECISION LINE AT 0.5 SO THAT:
  - P(X)'S ≥ 0.5 WILL BE CLASSIFIED AS 1
  - P(X)'S < 0.5 WILL BE CLASSIFIED AS 0

	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
$Y = 1$	16	95	111
$Y = 0$	27	35	62
Total	43	130	173

THIS IS A CONFUSION MATRIX:

- TELLS PERFORMANCE OF OUR ALGORITHM

HOW TO READ IT?

16 POINTS WERE CLASSIFIED AS 0 BY OUR MODEL BUT INSTEAD THEY WERE 1'S

35 POINTS THE OTHER WAY AROUND

WHILE 95 & 27 WERE CLASSIFIED CORRECTLY

- WE CAN ALSO RETRIEVE THE ACCURACY OF THE MODEL BY A RATIO:

$$\frac{\text{Correctly Classified}}{\text{Total Points}} = \frac{27 + 95}{27 + 95 + 16 + 35} = 0.705$$

THE DECISION BOUNDARY WAS 0.5, BY CHANGING IT WE'LL GET DIFFERENT RESULT BUT HOW DO WE CHOOSE THIS DECISION BOUNDARY?

Training Error vs Test Error

- What about quality of our model?
- Test error rate is one measure:

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

→ probability of making a wrong prediction!

- By computing this on training data → training error rate
- By computing this on new data → test error rate

Estimating the Regression Coefficients

- Params are estimated by maximum likelihood

Binomial likelihood:

$$L(\beta, y) = \prod_{i:y_i=1} p(x_i) \prod_{j:y_j=0} (1 - p(x_j))$$

Log likelihood:

$$L(\beta, y) = \sum_{i=1}^{n} (y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i)))$$

- What we wanna do is maximise likelihood:

Iteratively reweighted least squares method

- Model Summary output usually reports a Wald-Test for testing $\beta_j = 0$:

$$\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim \mathcal{N}(0,1)$$

The likelihood-ratio test is a better, more general, test for comparing models:

$$-2 \log \frac{L(\hat{\beta}_0)}{L(\hat{\beta}_1)} = -2 \left[\log L(\hat{\beta}_1) - \log L(\hat{\beta}_0)\right] \sim \chi^2_{df_1 - df_0}$$

Often called deviance tests as they compare deviance.

Diagnosis plots for model checkin are similar to linear regression.

Model Probability

- Since y is binary,

$$\mathbb{E}(Y_i = 1 \mid x) = 1 \cdot \mathrm{P}(Y_i = 1 \mid x) + 0 \cdot \mathrm{P}(Y_i = 0 \mid x) = \mathrm{P}(Y_i = 1 \mid x)$$

So what we see is that by modeling the probs, we also model the expectation of y.

Generalised Linear Models

$\mathbb{E}$ = expectation

- In linear regression, we model $\mathbb{E}(y|x)$ directly as linear combination
- In logistic regression, we model the logit of $\mathbb{E}(y|x)$ as linear combination
- We can then say:

Both are generalised linear models → where function of the mean is linear

$$g(\mathbb{E}(y|x)) = x\beta$$

Given that g = monotone function:

$$\mathbb{E}(y|x) = g^{-1}(x\beta)$$

Multinomial Logistic Regression

- Imagine y_i taking values among $1, \dots, K$
- Using one-hot encoding:

$y_i = (0, 1, 0, \dots, 0)$ — where the one stands for the right class

It is clear that y_i follows a multinomial distribution

It's probability vector is $(p_{i1}, \dots, p_{iK})$

So now we have $K-1$ probabilitiesss.

K = Big K

k = small k

- Generalization to multiple classes:
  - Select an arbitrary class as baseline (K here)
  - Consider for another class k the odds of being k rather than K

$$\log = \frac{P(y=K|X=x)}{P(y=k|X=x)} = x\beta^k$$

The params, $\beta^1, \dots, \beta^{K-1}$ are each a vector of length $p+1$

Probs are:

$$P(y=k|X=x) = \frac{e^{x\beta^k}}{1 + \sum_{c=1}^{K-1} e^{x\beta^c}}$$

This is for small k

$$P(y=K|X=x) = \frac{1}{1 + \sum_{c=1}^{K-1} e^{x\beta^c}}$$

This is for Big K

They sum to 1

- Most common choice = Predict that y belongs to class with highest prob:

$$\hat{y} = \underset{k}{\text{argmax}}\, \hat{P}(y=k|X=x)$$

- This is obviously not the only way to generalize several categories.