Classification problems

So far, we have considered regression problems: Given an X_0, predict the corresponding outcome Y_0. Often $Y_0 \in \mathbb{R}$.

In many applications, we are interested in predicting what *class* a data point belongs to:

1. Based on `Annual income` and `Monthly credit card balance`, predict whether an individual will `Default` on their credit card payment.
   - The outcome credit card `Default` is binary – an individual can belong to one of two classes (default or no default).

Classification problems

So far, we have considered regression problems: Given an X_0, predict the corresponding outcome Y_0. Often $Y_0 \in \mathbb{R}$.

In many applications, we are interested in predicting what *class* a data point belongs to:

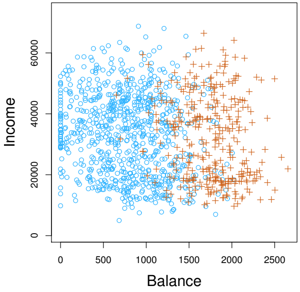
2. In the ER, a person with a certain set of symptoms could have one of three `medical condition`s: Stroke, drug overdose, or epileptic seizure. Which of the three conditions does the individual have?

    - The outcome `medical condition` has three categories/classes.

Classification: predicting a categorical outcome

So we want ML methods to *classify* data points.

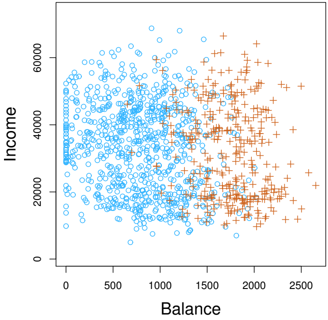
We call such a problem a *classification* problem.

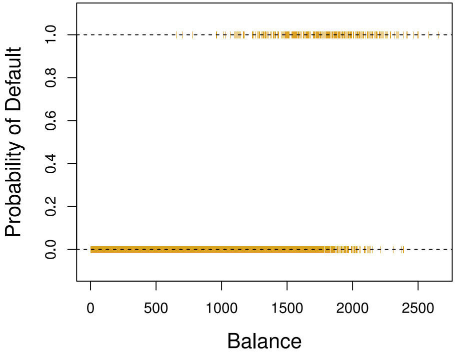


We want a model that guesses the color of the data points.

Logistic regression (binary classification)

A popular classification model is Logistic Regression.



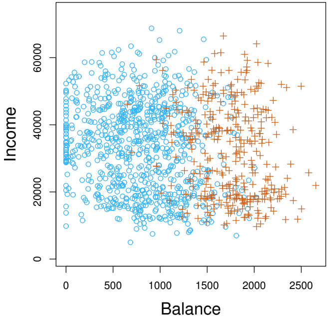


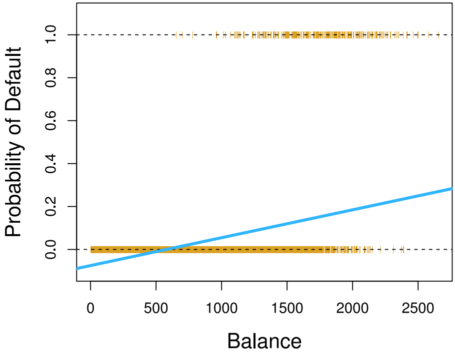
In Logistic Regression with 1 variable (here `Balance`), we assume that increasing the value of the variable monotonously increases the likelihood of one class.

Tempting to use Linear Regression. **Q:** Why should we not?

Logistic regression (binary classification)

A popular classification model is Logistic Regression.



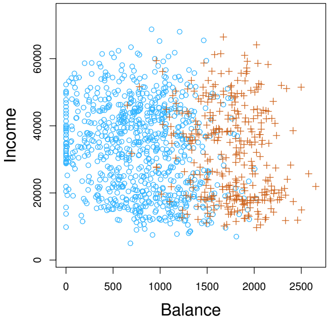


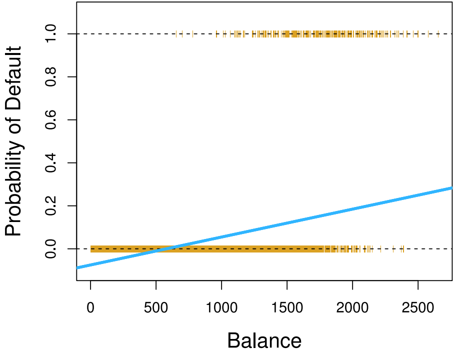
In Logistic Regression with 1 variable (here `Balance`), we assume that increasing the value of the variable monotonously increases the likelihood of one class.

Tempting to use Linear Regression. **Q:** Why should we not?

Logistic regression (binary classification)

A popular classification model is Logistic Regression.



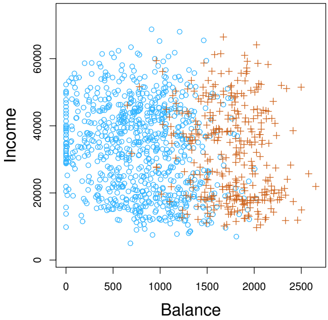


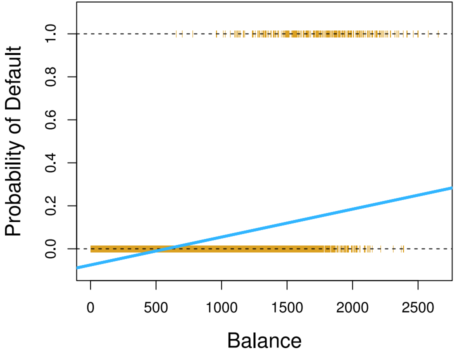
Problems with Linear Regression for modeling prob. of default $p(X)$:

- This could give us negative probabilities $p(X) < 0$
- This could give us probabilities $p(X) > 1$.

Logistic regression (binary classification)

A popular classification model is Logistic Regression.



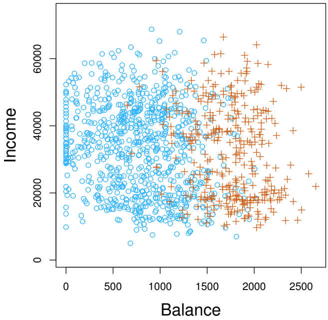


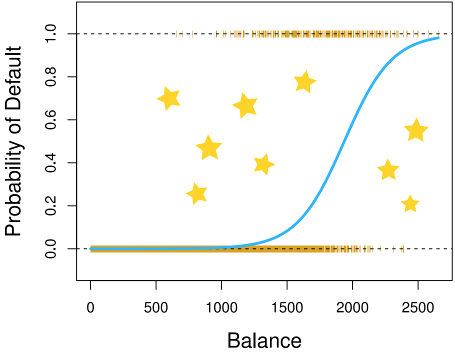
Instead, we seek a model for $p(X)$ that:

- Asympotically gives us $p(X) \to 1$ in one direction of X,
- Asymptotically gives us $p(X) \to 0$ in another direction of X.

Logistic regression (binary classification)

A popular classification model is Logistic Regression.

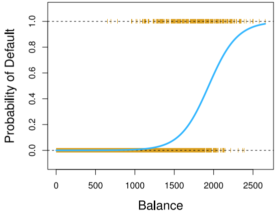




Instead, we seek a model for $p(X)$ that:

- Asympotically gives us $p(X) \to 1$ in one direction of X,
- Asymptotically gives us $p(X) \to 0$ in another direction of X.

The Logistic Function



What we are looking for is the Logistic Function

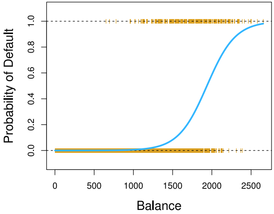
$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

(Often, we refer to this funtion as the sigmoid, $p(X) = \sigma(\beta_0 + \beta_1 X)$)

The exponent looks like our Linear Regression, but.. The Logistic Function has nice properties.

- What happens when $\beta_0 + \beta_1 X \to \infty$?
- What happens when $\beta_0 + \beta_1 X \to -\infty$?
- What happens when $\beta_0 + \beta_1 X = 0$?

Odds



The Logistic Function always produces an s-shaped curve. (great!)

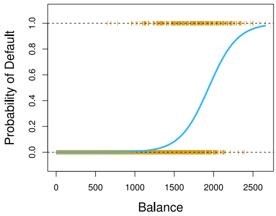
$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

Sometimes, we like to think about likelihoods, not in terms of $p(X)$, but instead,

$$\frac{p(X)}{1-p(X)}$$

Q: If $p(X)$ is the probability of a data point X corresponding to a `Default`. How can we interpret this fraction?

Odds



The Logistic Function always produces an s-shaped curve. (great!)

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

Sometimes, we like to think about likelihoods, not in terms of $p(X)$, but instead,

$$\text{odds:} \quad \frac{p(X)}{1-p(X)} = \frac{\text{Probability of default}}{\text{Probability of NOT default}}$$

Q: If $p(X)$ is the probability of a data point X corresponding to a default. How can we interpret this fraction?

Odds (examples)

Sometimes, we like to think about likelihoods, not in terms of odds

$$\text{odds:} \quad \frac{p(X)}{1 - p(X)}$$

Q: What are the odds if,

- Probability of heads is $p(X) = \frac{1}{2}$
- Probability of rain is $p(X) = \frac{3}{5}$
- Probability of canteen food being great is $p(X) = \frac{1}{3}$

Odds (examples)

Sometimes, we like to think about likelihoods, not in terms of odds

$$\text{odds:} \quad \frac{p(X)}{1-p(X)}$$

Q: What are the odds if,

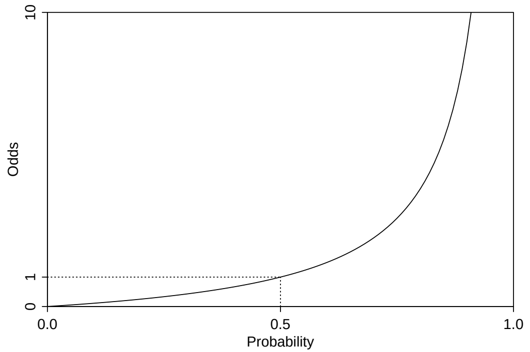
- Probability of heads is $p(X) = \frac{1}{2}$. Answer: 1.
- Probability of rain is $p(X) = \frac{3}{5}$. Answer: $\frac{3}{2}$.
- Probability of canteen food being great is $p(X) = \frac{1}{3}$. Answer: $\frac{1}{2}$.

Interpretation of odds of c/d: For every c times the event of interest occurs, it the event will not occur d times.

Probability and odds

Odds are positive real numbers.

The odds of an event A are greater than the odds of an event B exactly when the probability of A is greater than the probability of B.



Odds of 1 correspond to it being equally likely that the event happens or not (a probability of 0.5).

Odds ratio

The odds in two groups can be compared by their *odds ratio*.

Often, we use the odds ratio to compare the relative chance of an event happening under 2 different conditions.

An odds ratio of 5 means that the odds of having cancer is 5 times higher for patient who smokes than a patient who does not smoke.

Here, the conditions are “patient smokes” and “patient does not smoke.”

For the logistic function, the odds are

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

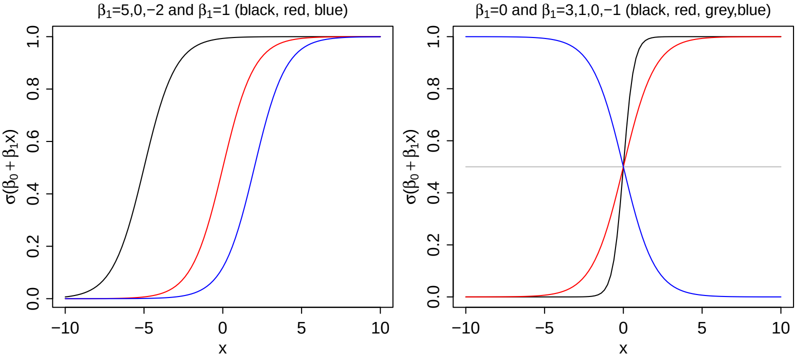
Taking the logarithm gives us the *log odds*, or *logit*,

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

Notice: The difference in two log-odds is the log of a odds ratio.
Notice: The logistic regression model has a logit that is linear in X.
($p(X)$ changes monotonously with X)

Probabilities as function of feature x

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



- $-\beta_0/\beta_1$ determines the point where $p(x) = 0.5$.
- β_1 determines steepness as $p'(x) = \beta_1 p(x)(1 - p(x))$.

Interpretation of coefficients

One unit change in feature x_1 means

- a change of β_1 in log-odds
- a multiplicative change in odds by a factor e^{β_1}
  - e^{β_1} is an *odds ratio*.
  - β_1 is a *log odds ratio*.

Because of the nonlinear translation between odds and probability, it is hard to communicate the change in probability, but

- The logit-transformation is monotone.
- Positive β_1 gives positive change in probability.
- A higher β_1 gives a steeper curve.

Logistic regression (multiple features)

Generalizing to the case of several variables, logistic regression models the conditional probability of $Y = 1$ given features $x_1, \ldots, x_p$ as a logistic function of a linear combination of the features:

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

(X includes the "constant feature" x_0)

Rearranging this gives a model for the *odds*

$$\frac{p(X)}{1 - p(X)} = e^{X\beta}$$

and a **linear model** for the *log-odds*!

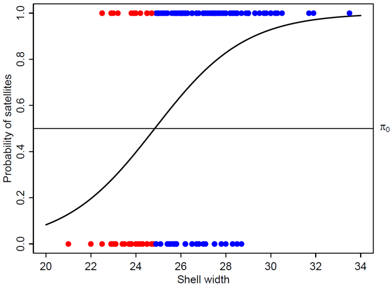
$$\log \frac{p(X)}{1 - p(X)} = X\beta$$

Prediction

To make a prediction $\hat{Y}$, we classify to the class with highest probability, i.e. that with $p(x) > 0.5$

$$\hat{Y}_i = \begin{cases} 1, & \mathrm{P}(Y_i = 1 \,|\, x_i) > 0.5 \\ 0, & \mathrm{P}(Y_i = 1 \,|\, x_i) \leq 0.5 \end{cases}$$

Prediction



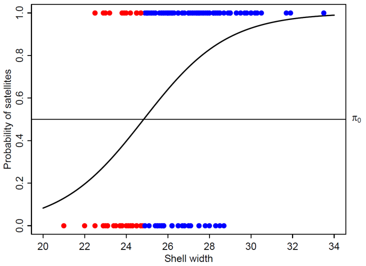
	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
$Y = 1$	16	$\boxed{95}$	111
$Y = 0$	$\boxed{27}$	35	62
Total	43	130	173

This is a **confusion matrix.** It tells us the performance of our classification algorithm.

$TP = 95$, $TN = 27$, $FP = 35$, $FN = 16$.

Accuracy: $(27 + 95)/173 = 0.705$. (The fraction we get right.)

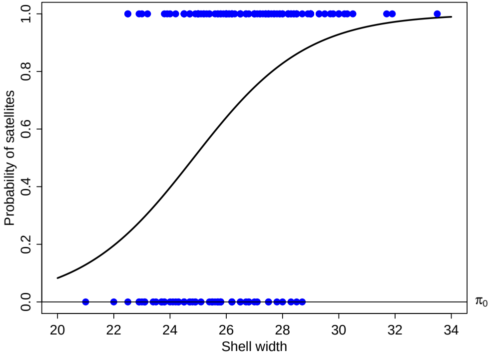
Decision boundaries and decision regions



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

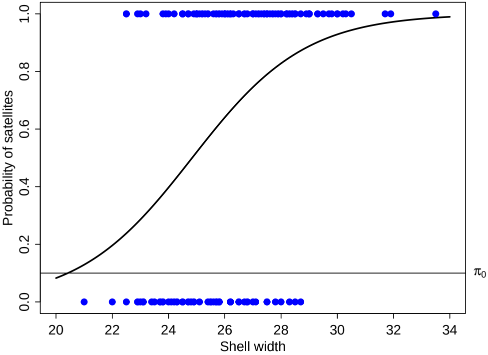
We chose to make the most-likely class our prediction. For this choice, the values of the features for which $p(x) = 0.5$ is the *decision boundary* – it splits the feature space into two *decision regions*: On "one side" we predict class 0, on the other class 1.
We could choose a different threshold than 0.5 for the probability - what would be the effect of that?

Changing the probability threshold



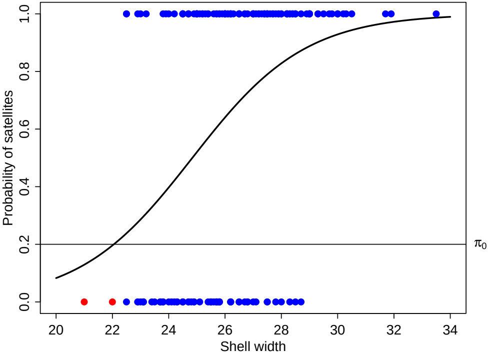
	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
$Y = 1$	0	$\boxed{111}$	111
$Y = 0$	$\boxed{0}$	62	62
Total	0	173	173

Changing the probability threshold



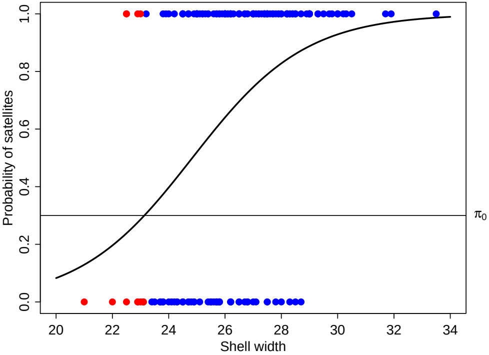
	$\hat{Y}=0$	$\hat{Y}=1$	Total
$Y=1$	0	$\boxed{111}$	111
$Y=0$	$\boxed{0}$	62	62
Total	0	173	173

Changing the probability threshold



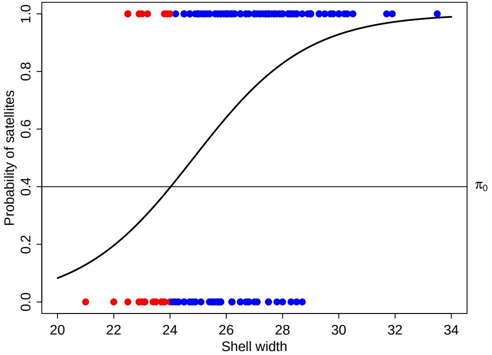
	$\hat{Y}=0$	$\hat{Y}=1$	Total
$Y=1$	0	111	111
$Y=0$	2	60	62
Total	2	171	173

Changing the probability threshold



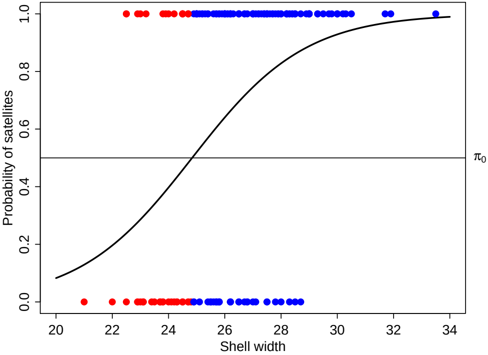
	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
$Y = 1$	4	$\boxed{107}$	111
$Y = 0$	$\boxed{9}$	53	62
Total	13	160	173

Changing the probability threshold



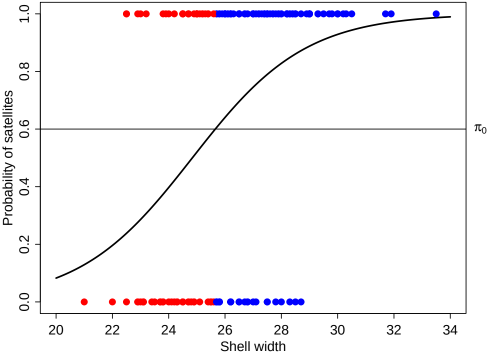
	$\hat{Y}=0$	$\hat{Y}=1$	Total
$Y=1$	8	$\boxed{103}$	111
$Y=0$	$\boxed{17}$	45	62
Total	25	148	173

Changing the probability threshold



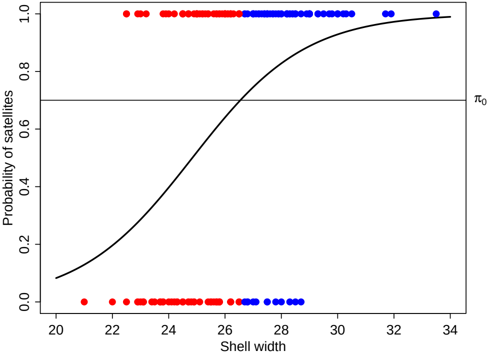
	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
$Y = 1$	16	$\boxed{95}$	111
$Y = 0$	$\boxed{27}$	35	62
Total	43	130	173

Changing the probability threshold



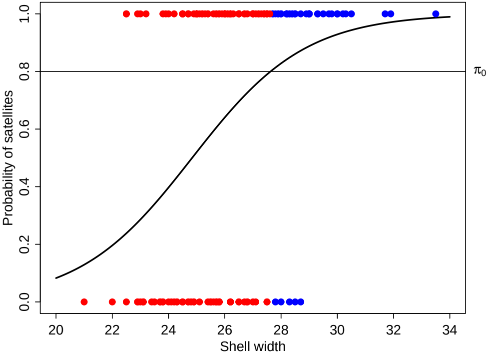
	$\hat{Y}=0$	$\hat{Y}=1$	Total
$Y=1$	30	$\boxed{81}$	111
$Y=0$	$\boxed{35}$	27	62
Total	65	108	173

Changing the probability threshold



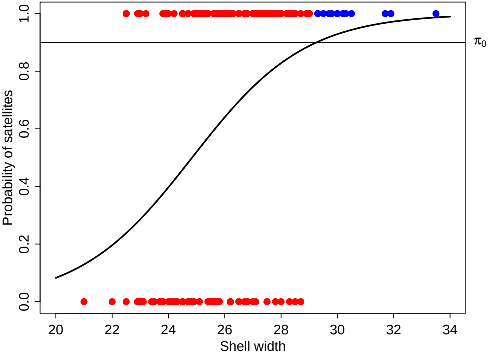
	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
$Y = 1$	52	$\boxed{59}$	111
$Y = 0$	$\boxed{50}$	12	62
Total	102	71	173

Changing the probability threshold



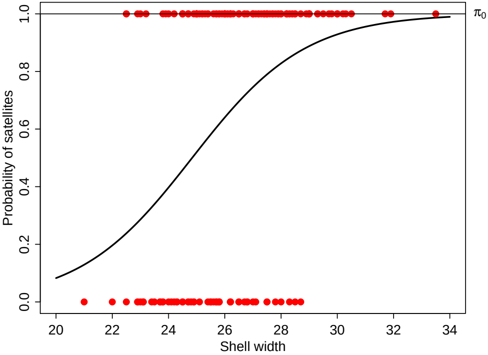
	$\hat{Y}=0$	$\hat{Y}=1$	Total
$Y=1$	71	$\boxed{40}$	111
$Y=0$	$\boxed{57}$	5	62
Total	128	45	173

Changing the probability threshold



	$\hat{Y}=0$	$\hat{Y}=1$	Total
$Y=1$	97	14	111
$Y=0$	62	0	62
Total	159	14	173

Changing the probability threshold



	$\hat{Y}=0$	$\hat{Y}=1$	Total
$Y=1$	111	$\boxed{0}$	111
$Y=0$	$\boxed{62}$	0	62
Total	173	0	173

Decision boundaries are linear

Predicting to the class with highest probability is the same as predicting to the class with highest log-odds.

Log-odds are linear and thus much easier to reason about.

Training error vs test error

How do we measure prediction quality?

Test error rate–the probability of making a wrong prediction–is one measure.

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

If we compute this on the training data, we call it the training error rate. If new data is used, it is called the test error rate.

We will discuss a plethora of performance measures for classification later.

Estimating the regression coefficients

Parameters are estimated by maximum likelihood.

The binomial likelihood function is

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

Remember that $p(x_i) = \sigma(x_i\beta)$ is a function of β.

The log likelihood:

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

The likelihood needs to be maximised numerically.implemented by the "iteratively reweighted least squares method".

Statistical tests and model checking

Model summary output usually reports a Wald-test for testing $\beta_j = 0$:

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

The likelihood-ratio test is a better, more general, test for comparing two nested models (with parameter vectors $\hat{\beta}_0$, $\hat{\beta}_1$):

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

For logistic regression, the test is often called deviance tests as they compare the deviance ($-2 \log L -$ constant) of two models.

Diagnostic plots for model checking are similar to linear regression, but a harder to interpret due to the discreteness of the response variable.

Model for probability is a model for expectation

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, in fact, by modelling the probability we also model the expectation of Y - just like we did in the Gaussian linear regression model.

Generalised linear models

In linear regression, we model $\mathbb{E}(Y \mid x)$ directly as a linear combination of predictors.

In logistic regression, we model the logit function of $\mathbb{E}(Y \mid x)$ as a linear combination of predictors.

Both are generalised linear models: Models where *a function of the mean* is linear:

$$g(\mathbb{E}(Y \mid X)) = X\beta$$

The *link function* g is always a monotone function, so

$$\mathbb{E}(Y \mid X) = g^{-1}(X\beta).$$

For Gaussian, the link function g is the identity function.
For logistic regression, the link function g is the logit function.

Other link functions can be used with the binomial distribution.

Multiclass classification

What if Y_i takes values among several classes $1, \ldots, K$?

Using the one-hot encoding, $Y_i = (0, 1, \ldots, 0)$ with a one in the entry corresponding to the group, it is clear that Y_i follows a multinomial distribution with probability vector $(p_{i,1}, \ldots, p_{i,K})$.

Now rather than having one probability (or one log-odds) to model, we have $K - 1$ of them!

Multinomial logistic regression

Here's how we usually generalize to multiple classes...

Select an arbitrary class (here K) as the *baseline* and consider for another class k the odds of being in class k rather than class K:

$$\log \frac{\mathrm{P}(Y = k \mid X = x)}{\mathrm{P}(Y = K \mid X = x)} = X\beta^k$$

Model parameters $\beta^1, \ldots, \beta^{K-1}$ are each a vector of length $p + 1$.

The probabilities are

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

and for class K

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

and they sum to one as they should!

Prediction - multiclass classification

The most common choice is to predict that Y belongs to the class with highest probability.

$$\hat{Y} = \arg\max_k \hat{\mathrm{P}}(Y = k \,|\, X = x)$$

In a couple of weeks you should know the rationale behind this choice.

Softmax regression

There are other ways to generalize to several categories...

Softmax regression uses an alternative coding with no baseline category, which only changes the interpretation of the coefficients.

$$\mathrm{P}(Y = k \mid X = x) = \frac{e^{X\beta^k}}{\sum_{c=1}^{K} e^{X\beta^c}}$$

Otherwise exactly the same as multinomial logistic regression before. Predictions are unchanged.