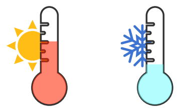
Classification I: Logistic Regression

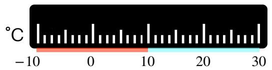
Lecture 4

Termeh Shafie

Regression vs. Classification



- What is the temperature going to be tomorrow?
- Will it be hot or cold tomorrow?



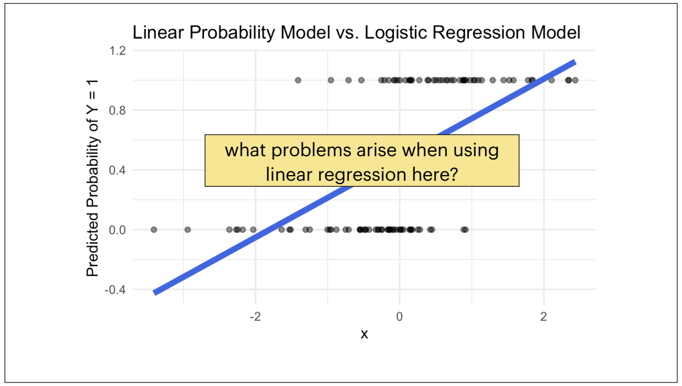
continuous variable
(from $-\infty$ to ∞)

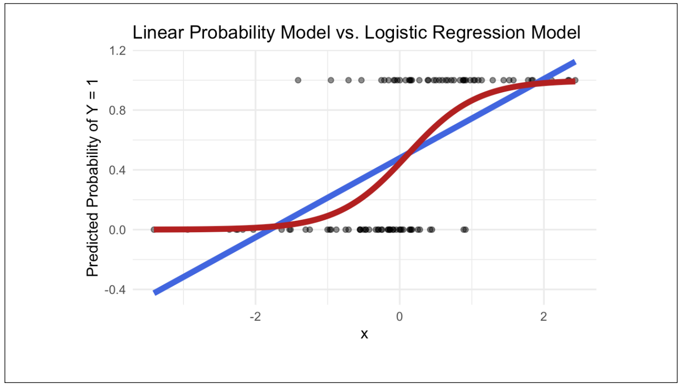
categorical variable
(0 to 1)

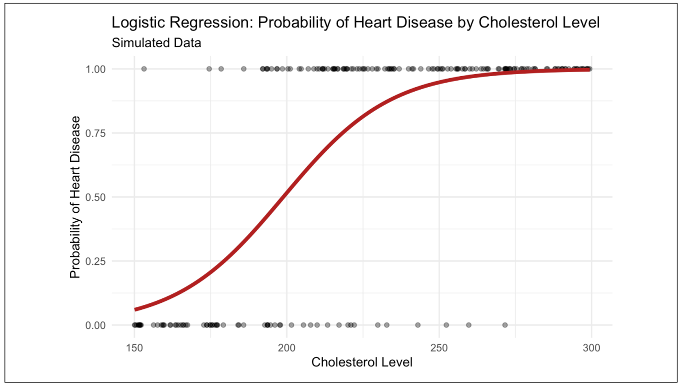
Linear Probability Model

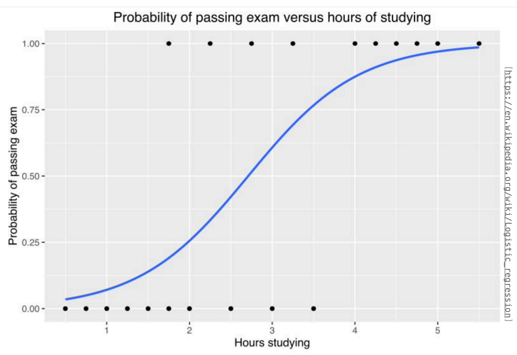
predictions

1	0.80
0	0.55
0	0.30
1	0.65
1	0.50


Linear Probability Model vs. Logistic Regression Model
what problems arise when using linear regression here?
Predicted Probability of Y = 1
x


Linear Probability Model vs. Logistic Regression Model
Predicted Probability of Y = 1
x


Logistic Regression: Probability of Heart Disease by Cholesterol Level
Simulated Data
Probability of Heart Disease
Cholesterol Level



Redefining The Response

original Y — 0, 1

Y as probability — $[0, 1]$

?

$Y' \in (-\infty, \infty)$ — $-\infty$ to ∞

how transform Y from $\{0,1\}$ to the real line?

Link Functions

a method to get "non-linear" linear regression
(more on this topic in a later lecture...)

$$y = X\beta$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$y = g^{-1}(X\beta)$$

the **link function** transforms the probabilities of the levels of a categorical response variable to a continuous scale that is unbounded

the **link function** transforms back the expectation of the response to the linear function

Logistic Regression

logit link function and log odds

$$y = X\beta \qquad\qquad y = g^{-1}(X\beta)$$

$$\underbrace{\log\underbrace{\left(\frac{p}{1-p}\right)}_{\text{odds}}}_{\text{logit link function}} = \beta_0 + \beta_1 x_1$$

$$p = \frac{e^{\beta_0+\beta_1x_1}}{1+e^{\beta_0+\beta_1x_1}} = \text{[a little algebra]} = \frac{1}{1+e^{-(\beta_0+\beta_1x_1)}}$$

Redefining The Response

original Y

$\downarrow$

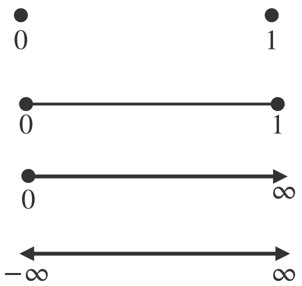
Y as probability

$\downarrow$

odds of Y

$\downarrow$

$Y' \in (-\infty, \infty)$



Logistic Regression

$$y = X\beta$$

our link function is

$$g(x) = \log\frac{x}{1-x}$$

which has the inverse

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

$$\boxed{y = g^{-1}(X\beta)}$$ general case

$$\boxed{p = \frac{e^{X\beta}}{1+e^{X\beta}}}$$ specific case

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\beta = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\beta = \mu$	$\mu = \mathbf{X}\beta$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\beta = -\mu^{-1}$	$\mu = -(\mathbf{X}\beta)^{-1}$
Gamma					
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\beta = \mu^{-2}$	$\mu = (\mathbf{X}\beta)^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\beta = \ln(\mu)$	$\mu = \exp(\mathbf{X}\beta)$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\beta = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\beta)}{1+\exp(\mathbf{X}\beta)} = \frac{1}{1+\exp(-\mathbf{X}\beta)}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\beta = \ln\left(\frac{\mu}{n-\mu}\right)$	
Categorical	integer: $[0, K)$ *K*-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1	outcome of single *K*-way occurrence		$\mathbf{X}\beta = \ln\left(\frac{\mu}{1-\mu}\right)$	
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1, ..., *K*) out of *N* total *K*-way occurrences			

https://en.wikipedia.org/wiki/Generalized_linear_model#Link_function

Interpreting Logistic Regression Models

- we want to create a spam filter based on 3921 observations/emails
- simple model, one predictor: '`to_multiple`'

```
Call:
glm(formula = spam ~ to_multiple, family = binomial, data = email)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.477  -0.477  -0.477  -0.477   2.809

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.11609    0.05618 -37.665  < 2e-16 ***
to_multipleyes -1.80918    0.29685  -6.095  1.1e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2437.2  on 3920  degrees of freedom
Residual deviance: 2372.0  on 3919  degrees of freedom
AIC: 2376

Number of Fisher Scoring iterations: 6
```

Interpreting Coefficients

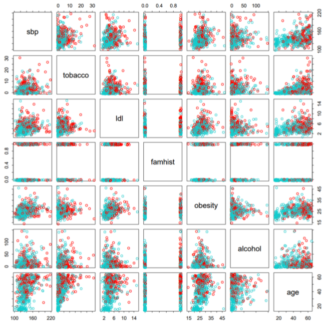
Probability p	**Odds** p/(1-p)	**Log Odds** log[p/(1-p)]
0.1	0.1111	-2.1972
0.5	1	0
0.9	9	2.1972

Example: South African Heart Disease

- From Western Cape,South Africa in early 80s
- Coronary Risk Factor Study (CORIS)
- High incidence of myocardial infarction (MI) in region: 5.1%
- Measurements on seven predictors (risk factors)
- 160 cases, 302 controls. Ages 15-64.
- Outcome is presence/absence of MI at time of survey
- Goal:
  - to identify relative strengths and directions of risk factors
  - intervention study aimed at educating the public on healthier diets

[For more info see ESL 4.4.2]

Example: South African Heart Disease

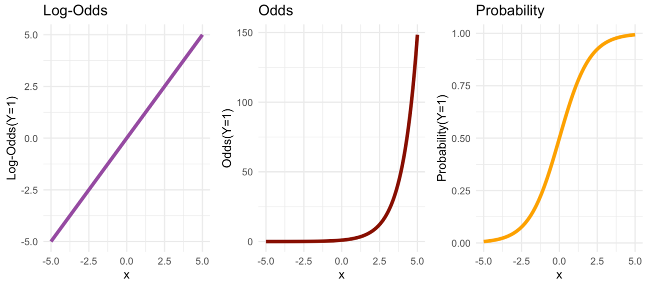


cases = red
controls = turquoise

Example: South African Heart Disease

term	estimate	std.error	statistic	p.value
(Intercept)	-4.130	0.964	-4.283	0.000
sbp	0.006	0.006	1.023	0.306
tobacco	0.080	0.026	3.034	0.002
ldl	0.185	0.057	3.219	0.001
famhistPresent	0.939	0.225	4.177	0.000
obesity	-0.035	0.029	-1.187	0.235
alcohol	0.001	0.004	0.136	0.892
age	0.043	0.010	4.181	0.000

Interpreting Coefficients



Estimating Coefficients: MLE

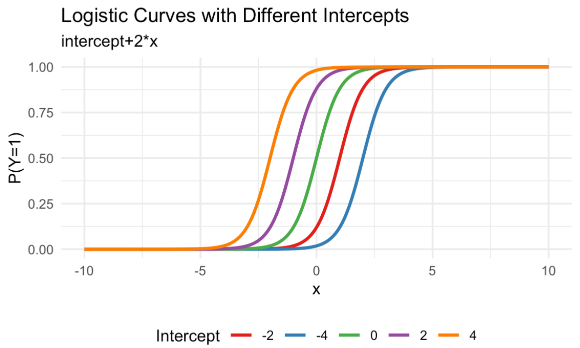
$$\prod_{i;y_i=1} p(x_i) \qquad \prod_{i;y_i=0} 1 - p(x_i)$$

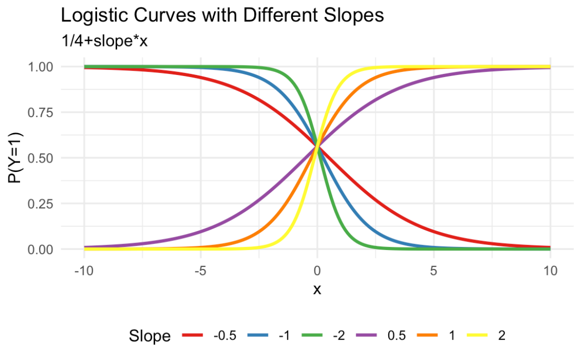
$$L(\beta_0, \beta_1) = \prod_{i;y_i=1} p(x_i) \cdot \prod_{i;y_i=0} 1 - p(x_i)$$

$$L(\beta_0, \beta_1) = \prod_{i=1} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

$$l(\beta_0, \beta_1) = \sum_{i=1} y_i \log(p(x_i)) + (1 - y_i)\log(1 - p(x_i))$$

[full proof:https://arunaddagatla.medium.com/maximum-likelihood-estimation-in-logistic-regression-f86ff1627b67

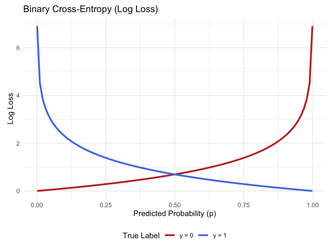




Binary Cross Entropy

$$l(\beta_0, \beta_1) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(p(x_i)) + (1 - y_i)\log(1 - p(x_i))$$

Loss function



Assessing Model Performance

- Did it make the correct prediction?
  - accuracy
  - sensitivity
  - specificity
- How well does to perform in distinguishing classes correctly?

Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (TN)
	Negative	False Positive (FP)	True Negative (TN)

Confusion Matrix: Accuracy

correct predictions

True Positive (TP) + True Negative (TN)

---

all predictions

False Negative (TN) + False Positive (FP) +

True Positive (TP) + True Negative (TN)

How often is the model correct?

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (TN)
	Negative	False Positive (FP)	True Negative (TN)

Confusion Matrix: Sensitivity/Recall

correctly predicted positives

True Positive (TP)

---

actual positives

False Negative (TN) + True Positive (TP)

How often is the model correct for Positive Cases?

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (TN)
	Negative	False Positive (FP)	True Negative (TN)

Confusion Matrix: Specificity

$$\frac{\text{correctly predicted negatives: True Negative (TN)}}{\text{actual negatives: True Negative (TN)} + \text{False Positive (FP)}}$$

How often is the model correct for Negative Cases?

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (TN)
	Negative	False Positive (FP)	True Negative (TN)

Confusion Matrix: Precision

$$\frac{\text{correctly predicted positives: True Positive (TP)}}{\text{all positives: False Positive (FP)} + \text{True Positive (TP)}}$$

How many of the predicted Positives are correct?

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (TN)
	Negative	False Positive (FP)	True Negative (TN)

Confusion Matrix: F1 Score

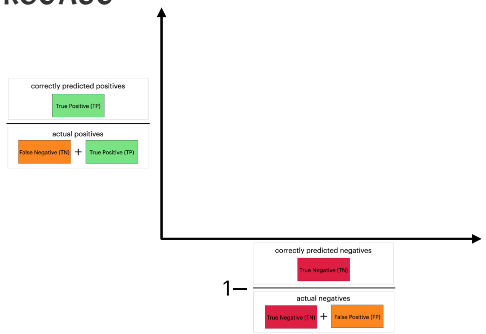
$$\frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

$$= \frac{2 \times precision \times recall}{precision + recall}$$

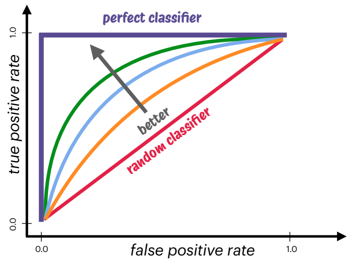
Combination of Precision (how often predicted positives ARE positive) and Recall (how often we correctly predict actual positives)

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (TN)
	Negative	False Positive (FP)	True Negative (TN)

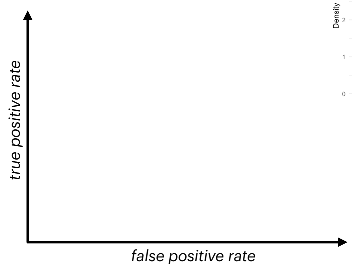
ROC AUC

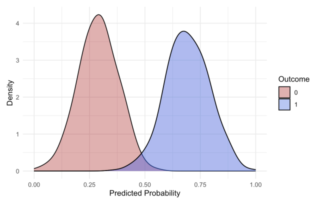


ROC AUC

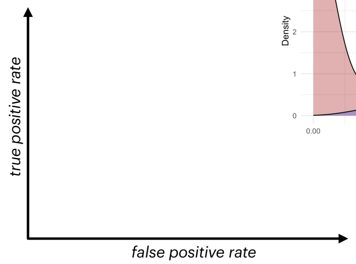


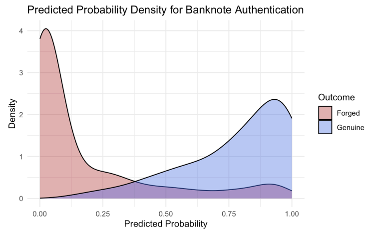
ROC AUC





ROC AUC





more realistic...

This Week's Practical

Logistic Regression: The Stock Market Data