Logistic Regression - Data Analysis and Visualization

Dr. Farshid Alizadeh-Shabdiz

Spring 2024

Classification

Classification

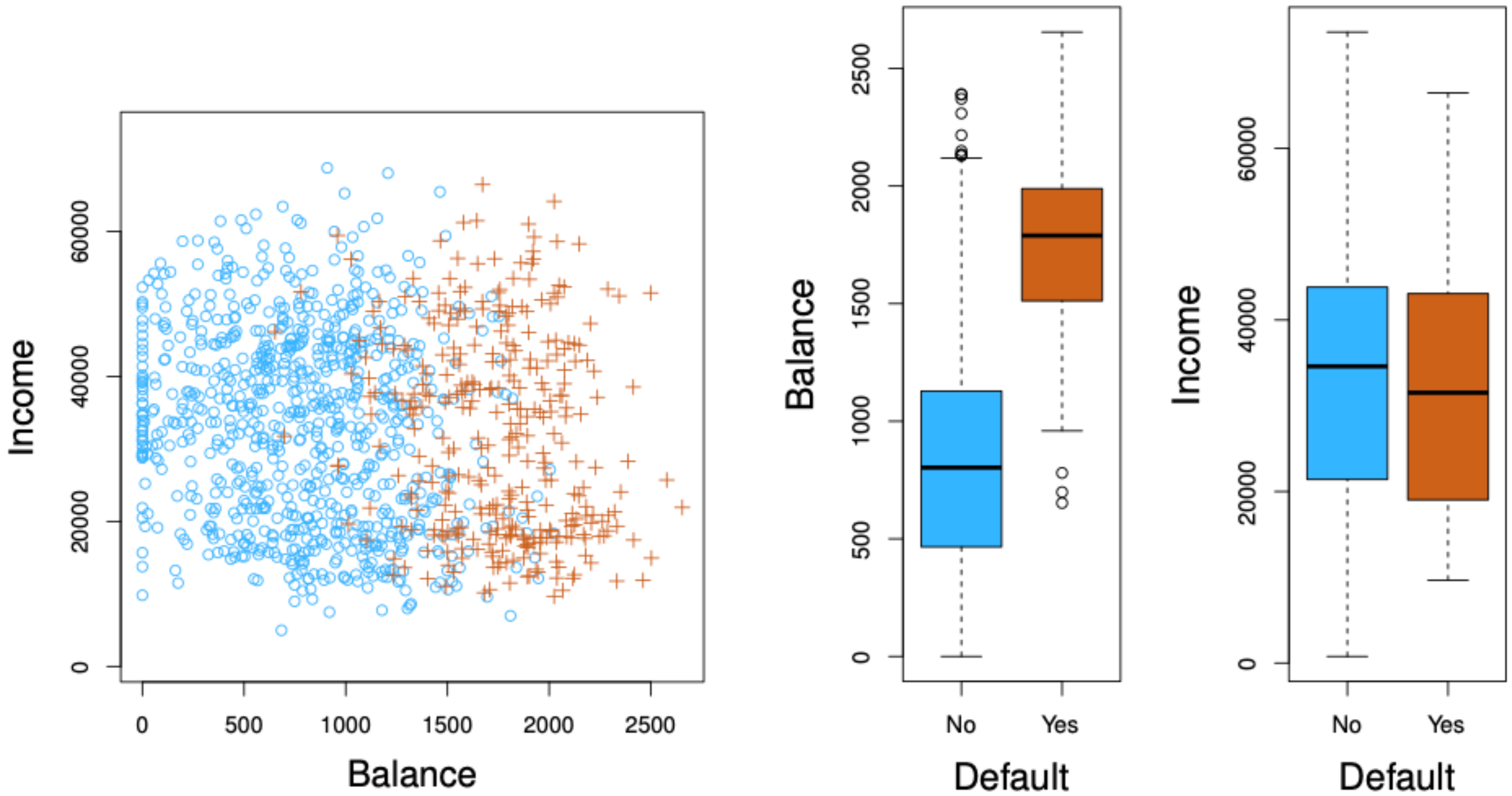
Classification is

- Predicting a qualitative output from
- A set of quantitative inputs/parameters, X
- And most of the time, also estimating quality of the estimate

For example

- Detecting spam email
- Detecting credit card fraud
- Diagnosing a patient's illness
- Object detection
- Face recognition
- News article classification
- Recommending products to customers

Example of Classification



Ref: In-depth introduction to machine learning, by T. Hastie and R. Tibshirani

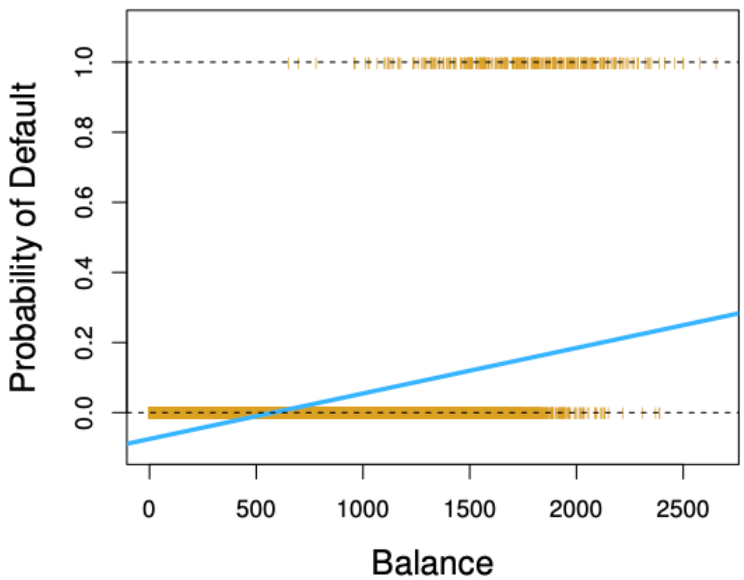
Linear Regression as classification

- Can linear regression solve classification with output

$$y = \begin{cases} 0 \\ 1 \end{cases} \ \& \ Threshold = \ 0.5$$

- Short answer is YES – linear regression does a good job, BUT
  - Linear regression can produce an output outside [0,1] !!
  - It can give a probability much higher than one for class-1
  - Also a probability much lower than zero for class-0
  - Many assumption of linear regression are not met

Linear Regression



Ref: In-depth introduction to machine learning, by T. Hastie and R. Tibshirani

Why Logistic Function

- The goal is
  - Finding probability of a success event, which is p (Diagnose is C).
  - Using linear combination of parameters to estimate p
- Note p changes between *[0,1]*, but linear combination of parameters change between $(-\infty, \infty)$ – Big discrepancy

- So probability of no-success is (1-p)
- The "odds ratio" is defined as $(\frac{p}{1-p})$, which mean how odd is having a success
- The odds ratio varies between $[0, \infty)$, closer to linear output
- Easy way to map a range of real positive to real numbers is *log* function
- So, we have it!

Logistic Regression

- So the output *y* will be found as

$$y = \frac{e^{\beta 0 + \beta 1 X 1}}{1 + e^{\beta 0 + \beta 1 X 1}}$$

- e (~2.71828) is constant. It is called Euler's number or "natural number"
- Note $(0 \leq y \leq 1)$

Another way of looking at logistic Function

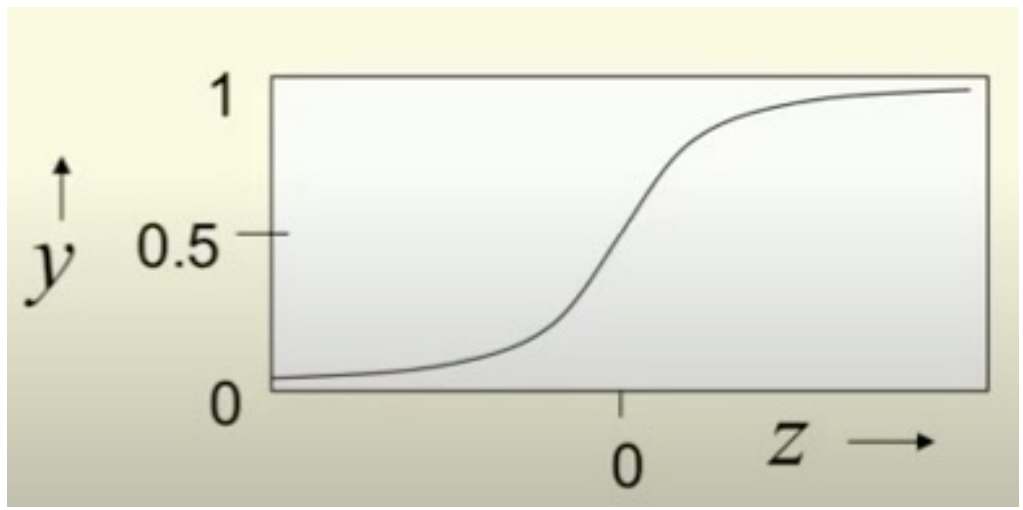
$$Z = \beta_0 + \beta_1 X_1$$

$$y = \frac{e^Z}{1+ez}$$

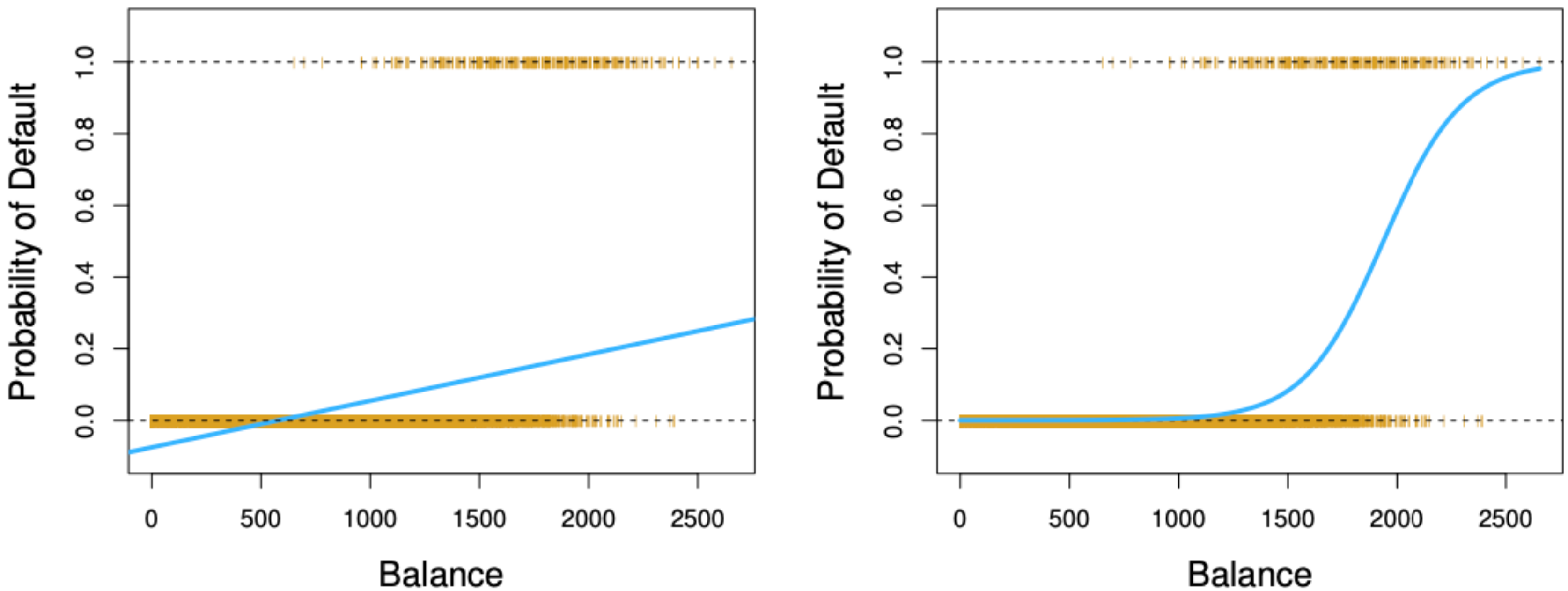
Logistic Function Is a Good Fit

- Logistic Function:

$$y = \frac{e^{Z}}{1+ez}$$



Linear vs Logistic Regression



Ref: In-depth introduction to machine learning, by T. Hastie and R. Tibshirani

- The above example is a yes and no answers. Orange dots are marking the answers.

Example – Prediction of Default

- Predicting default as a function of balance

```
> glm(default ~ balance , data=Default , family=binomial)

Call:  glm(formula = default ~ balance, family = binomial, data = Default)

Coefficients:
(Intercept)      balance
 -10.651331     0.005499

Degrees of Freedom: 9999 Total (i.e. Null);  9998 Residual
Null Deviance:      2921
Residual Deviance: 1596        AIC: 1600
```

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

Example – Prediction of Default Examples

- Probability of a person with $1000 balance default

$$y = \frac{e^{\beta 0 + \beta 1 X}}{1 + e^{\beta 0 + \beta 1 X}} = \frac{e^{-10.6 + 0.0055 \times 1000}}{1 + e^{-10.6 + 0.0055 \times 1000}} = 0.006$$

What if balance is $2000

$$y = \frac{e^{\beta 0 + \beta 1 X}}{1 + e^{\beta 0 + \beta 1 X}} = \frac{e^{-10.6 + 0.0055 \times 2000}}{1 + e^{-10.6 + 0.0055 \times 2000}} = 0.586$$

Multivariable Logistic Regression

- So the output y will be found as

$$y = \frac{exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}{1 + exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}$$

Note that $(\mathbf{0} \leq y \leq \mathbf{1})$

And

$$y = \begin{cases} 0, & Default \\ 1, & No\ default \end{cases}$$

Example –Default for Students

- What about probability of default for students

```
> glm(default ~ student , data=Default , family=binomial)

Call:  glm(formula = default ~ student, family = binomial, data = Default)

Coefficients:
(Intercept)   studentYes
    -3.5041       0.4049

Degrees of Freedom: 9999 Total (i.e. Null);  9998 Residual
Null Deviance:      2921
Residual Deviance: 2909        AIC: 2913
```

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

- Predicting default as a function of “Student” being an student.

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

Example – Default vs Several Variables

- Predicting default as a function of “Student”, income, and balance is as follows.

```
> glm(default ~ . , data=Default , family=binomial)

Call:  glm(formula = default ~ ., family = binomial, data = Default)

Coefficients:
(Intercept)   studentYes      balance       income
 -1.087e+01   -6.468e-01    5.737e-03    3.033e-06

Degrees of Freedom: 9999 Total (i.e. Null);  9996 Residual
Null Deviance:      2921
Residual Deviance: 1572        AIC: 1580
```

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

Multivariable Logistic Regression

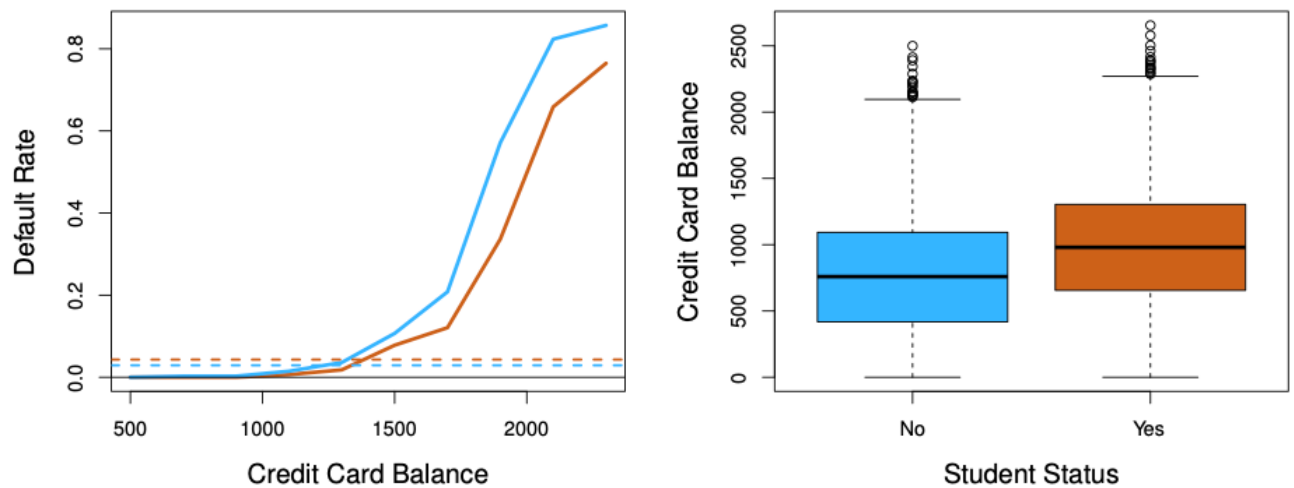
- Why student became negative here ?!

Note: correlation between variables in multi variable logistic regression can make inference hard.

```
> glm(default ~ . , data=Default , family=binomial)

Call:  glm(formula = default ~ ., family = binomial, data = Default)

Coefficients:
(Intercept)   studentYes      balance       income
 -1.087e+01   -6.468e-01    5.737e-03    3.033e-06

Degrees of Freedom: 9999 Total (i.e. Null);  9996 Residual
Null Deviance:      2921
Residual Deviance: 1572        AIC: 1580
```

Logic of the Results for Students

- Students have higher balance, so it is more likely for students to default
- But for a given balance, students default is lower



Ref: In-depth introduction to machine learning, by T. Hastie and R. Tibshirani

Interesting Math about Logistic Regression

- Logistic function has interesting mathematical properties

1. $\ln\left(\frac{y}{1-y}\right) = \beta_0 + \beta_1 X_1 + \cdots + BnXn$

   \- This is called *log odds* or *logit* transformation of *y*

2. $\frac{\partial y}{\partial z} = y(1-y)$ - derivative is important for numerical solutions

Odd ratio of Logistic Regression

$$\frac{y}{1-y} = \exp(\beta_0 + \beta_1 X_1 + \cdots + BnXn)$$

Interpretation

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}$$

- ▷ The regression equation can be used to predict risk of the event.
- ▷ The interpretation of the regression coefficient(s) are generally based on odds ratios.

Consider the odds ratio of an event for a given value of $x = x_a$ versus a given value of $x = x_b$.

- ▷ The estimated odds for a given value of $x = x_a$ is given by $\widehat{\text{odds}}_a = e^{\hat{\beta}_0 + \hat{\beta}_1 x_a}$
- ▷ The estimated odds for a given value of $x = x_b$ is given by $\widehat{\text{odds}}_b = e^{\hat{\beta}_0 + \hat{\beta}_1 x_b}$

The odds ratio then is given by

$$\widehat{OR}_{x_a \text{ versus } x_b} = \frac{\widehat{\text{odds}}_a}{\widehat{\text{odds}}_b} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_a}}{e^{\hat{\beta}_0 + \hat{\beta}_1 x_b}}$$

Interpretation

The odds of the event are $e^{\hat{\beta}_1(x_a - x_b)}$ higher for every $x_a - x_b$ unit increase in x.

Interpretation depends only on the difference in x values as opposed to their actual values.

$$
\begin{aligned}
\widehat{OR}_{x_a \text{ versus } x_b} &= \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_a}}{e^{\hat{\beta}_0 + \hat{\beta}_1 x_b}} \\
&= e^{(\hat{\beta}_0 + \hat{\beta}_1 x_a) - (\hat{\beta}_0 + \hat{\beta}_1 x_b)} \\
&= e^{\hat{\beta}_0 + \hat{\beta}_1 x_a - \hat{\beta}_0 - \hat{\beta}_1 x_b} \\
&= e^{\hat{\beta}_1 x_a - \hat{\beta}_1 x_b} \\
&= e^{\hat{\beta}_1 (x_a - x_b)}
\end{aligned}
$$

Confidence Interval

Confidence intervals for the logistic regression setting are based on the odds ratio.

The two-sided $100\% \times (1 - \alpha)$ confidence interval for $\widehat{OR}_{x_a \text{ versus } x_b}$ is:

$$e^{\left(\hat{\beta}_1 \pm z_{\frac{\alpha}{2}} \cdot SE_{\hat{\beta}_1}\right)(x_a - x_b)}$$

Where

- $SE_{\hat{\beta}_1}$ is the standard error of the regression coefficient and
- $z_{\frac{\alpha}{2}}$ is the value from the standard normal distribution with a right tail probability of $\alpha/2$.

An Example: Logistic Regression

We are interested in the association between cholesterol levels and having a coronary event in a high-risk patient population (who have had an event in the past). We collect cholesterol data for 50 subjects and then follow each for a year to see if they have another coronary event.

Explanatory variable is cholesterol level and our outcome is whether or not the subject had another coronary event.

Given the nature of our response variable, we perform a logistic regression. A summary of the beta estimates from the model are shown below.

Parameter	Estimate	Standard Error	p-value
β_0	-3.3848	0.5838	< 0.00001
β_1	0.1253	0.0362	0.0005

An Example: Logistic Regression

Parameter	Estimate	Standard Error	p-value
β_0	-3.725	1.753	0.0336
β_1	0.024	0.012	0.0420

Use these results to

- **predict the risk of another coronary event** for a high risk patient with a cholesterol level of 190.
- **calculate the odds ratio** for a coronary event of a high-risk patient with a cholesterol level of 190 versus a patient with a cholesterol level of 180.
- **calculate 95% confidence interval** for the odds ratio of having a coronary event for a patient with a cholesterol level of 190 versus a patient with a cholesterol level of 180.

An Example: Logistic Regression

Parameter	Estimate	Standard Error	p-value
β_0	-3.725	1.753	0.0336
β_1	0.024	0.012	0.0420

The risk of having a coronary event for a patient with a cholesterol level of 190 is predicted by :

$$\hat{p} = \frac{e^{-3.725+0.024*190}}{1+e^{-3.725+0.024*190}} = \frac{e^{0.835}}{1+e^{0.835}} = 0.697$$

An Example: Logistic Regression

The odds ratio of having a coronary event for a patient with a cholesterol level of 190 versus a patient with a cholesterol level of 180 is

$$\hat{OR}_{X_a versus X_b} = e^{\hat{\beta}_1 (X_a - X_b)} = e^{0.024*(190-180)} = e^{0.24} = 1.27$$

Interpretations:

- *The odds of having a coronary event are 1.27 times higher for every 10 Unit increase in cholesterol level.*
- *The odds ratio comparing any two individuals with cholesterol levels which are 10 units apart are the same.*

The quantity $e^{\hat{\beta}_1}$ is the odd ratio of the event for two individuals with x values that are 1 unit apart.

In other words, $e^{\hat{\beta}_1}$ is the relative increase in odds for every 1 unit increase in x.

An Example: Logistic Regression

The 95% confidence interval for the odds ratio of having a coronary event for a patient with a cholesterol level of 190 versus a patient with a cholesterol level of 180 is:

$$e^{\left(\hat{\beta}_1 \pm z_{\frac{\alpha}{2}} \cdot SE_{\hat{\beta}_1}\right)(x_a - x_b)} = e^{(0.024 \pm 1.96 \cdot 0.0122) \cdot 10}$$

$$= (1.004, 1.608)$$

We are 95% confident that the odds of having a coronary event are between 1.004 and 1.608 times higher for every 10-unit increase in cholesterol level.

Logistic Regression & Case Control Sampling

- When Logistic regression is used for rare events, the model will be trained for a none proportional ratio of samples
  - Like training with a set with 40% of rare event sample
- As a result the model will calculate probabilities wrong!

<u>Solution:</u>

- Case control sampling
  - Regression parameters $\boldsymbol{\beta_i}$ are accurate, and only the intercept $\boldsymbol{\beta_0}$ is not, which gets corrected by

$$\beta_0{}^* = \beta_0 + \log\left(\frac{p_{rare}}{1 - prare}\right) - \log\left(\frac{p_{set}}{1 - pset}\right)$$

- P_{rare}: actual probability of the rare event
- P_{set}: probability of the rare event in the training set

Control vs Case Sample Size

- Control to case ratio: In order to have smaller variance in the coefficients it is good to have more control samples.
- Question is how much more?
  - Rule of thumb is five to six times is sufficient

What about Multi Classification

- For example
  - Classifying news articles to sport, politics, family, kids, etc.
  - Classifying different people in a picture

Multiclass Logistic Regression or Multinomial Regression

- Logistic regression can be easily extended to more than two class prediction.

$$\Pr(y = k|X) = \frac{e^{\beta 0k + \beta 1k X1 + \cdots + \beta nk Xn}}{\sum_{j=1}^{K} e^{\beta 0j + \beta 1j X1 + \cdots + \beta nj Xn}}$$

K : capital K is total number of classes

k: small K is one of the classes

- Select the class with the highest probability

- This is also called “softmax” function
- Note - Linear regression cannot solve this problem

Generalized linear models (GLMs)[2]

- We can use ANY stochastic process for generating the error, not just normal distribution.
- GLM is an extension of linear regression that allows errors to be generated by a wide variety of distributions.
- In particular, any distributions in the "Exponential Family"
- GLMs extend the linear modeling capability of R to scenarios that involve non-normal error distributions. The idea is to obtain linear functions of the predictor variables by transforming the right side of the equation by a link function.

Error Family	**Link**	**Inverse of link**	**Used for**
Gaussian	identity	1	normally error
Poisson	log	exp	counts
Binomial	logit	$1/(1 + 1/\exp(x))$	proportions
Gamma	inverse	$1/x$	non-constant error

[2] https://en.wikipedia.org/wiki/Generalized_linear_model

Logistic Regression Function in R

- Logistic regression function in R provides
  - Coefficients of the model
  - p-value of the parameters same as linear regression
  - Also, “Z-statistics”, which is coefficient for normalized parameters

```
> tmp = glm(default ~ . , data=Default , family=binomial)
> summary(tmp)

Call:
glm(formula = default ~ ., family = binomial, data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

R Commands: Generalized Linear Models (GLMs)

Use the glm() function with binomial option

```
> glm(data$event~data$explanatory1 + data$explanatory2 + ... , family=
    binomial)
```

- Ensure that your event is coded as $1 = Event$ and $0 = non - event$ (numeric, not a factor variable)
- If one of the variables in the model is a factor variable, it is best to create dummy variables (1/0) so that you know exactly what the reference group is
- **"family"** parameter is a simple way of specifying a choice of variance and link functions When family is set to binomial, it tells R to perform logistic regression.

http://plantecology.syr.edu/fridley/bio793/glm.html

R commands: Generalized Linear Models (GLMs)

```
> glm(data$event~data$explanatory1 + data$explanatory2 + ... , family=
   binomial)
```

- Use the **summary() function** on the saved regression result to get regression equation and associated rests for each regression coefficient
- Use the **exp() function**, which computes the exponential value of a number e^x, on the resulting coefficients to obtain odds ratios for each regression coefficient
- Use the **predict() function** on the saved regression result to get the predicted risks for each observations

Logistic Regression: R commands

```
> data <- read.csv('cevent.csv')
# Simple logistic regression
> m <- glm(data$event ~ data$chol, family=binomial)
> summary(m)
Call:
glm(formula = data$event ~ data$chol, family = binomial)

Deviance Residuals:
    Min        1Q    Median        3Q       Max
-1.5752   -0.9629   -0.7217    1.1418    2.1732

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.72518    1.75307  -2.125   0.0336 *
data$chol    0.02359    0.01160   2.034   0.0420 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
             1
...
```

Logistic Regression In R

- Install.packages("ISLR")
- library(ISLR)
- Function
  - glm(y ~ X, data=DataName ,family = binomial)
  - glm(y ~ X+Z, data=DataName ,family = binomial)
  - glm(y ~ ., data=DataName ,family = binomial)
- Load data "Default"
  - Default as a function of Balance
  - Default as a function of Student
    - Default$studentBinFlag = ifelse(student=="Yes", 1, 0)
  - Default as a function of everything

Important note:
Logistic Regression is not Stable for fully separable classes.
In this case, other classification methods like Discriminant analysis has to be used.

Formal Inference in Logistic Regression

Formal Inference in Simple Logistic Regression

Formal inference in Simple Logistic Regression uses estimates of $\hat{\beta}1$.

$\boldsymbol{H_0 : \beta_1 = 0}$ (H_0 : there is no association between x and odds of the outcome)

$\boldsymbol{H_1 : \beta_1 \neq 0}$ (H_1 : there is an association between x and odds of the outcome)

$\boldsymbol{\beta_1 = 0}$ is equivalent to that the regression line had a slope of 0 (it would be a horizontal line), $\boldsymbol{\beta_1 = 0}$ means $\boldsymbol{OR = e^{\beta_1} = 1}$.

- The null hypothesis $\boldsymbol{\beta_1 = 0}$ is equivalent to the test of the odds ratio for a 1 unit increase in x being equal to 1 $\boldsymbol{(H_0 : OR = 1)}$.
- $\boldsymbol{H_0 : \beta_1 = 0}$ **or** $\boldsymbol{OR = 1}$ is rejected if $\hat{\beta}_1$ is sufficiently far from 0.
- We reject the claim that the population parameter β_1 is equal to 0 if $\hat{\beta}_1$, the sample statistic, is far from 0.

An Example: logistic regression inference

Formally test whether or not cholesterol is associated with risk of a coronary event at the $\alpha = 0.05$ level.

1. Set up the hypotheses and select the alpha level $H_0 : \beta_1 = 0$ or $OR = 1$ (there is no association between cholesterol levels and risk for a coronary event)

$H_1 : \beta_1 \neq 0$ or $OR 1$ (there is an association between cholesterol levels and risk for a coronary event) $\alpha = 0.05$

2. Select the appropriate test statistic

$z = \frac{\beta_1}{SE_{\beta_1}}$

3. State the decision rule

Determine the appropriate value from the standard normal distribution associated with a right hand tail probability of $\alpha/2 = 0.05/2 = 0.025$

$z_{\frac{\alpha}{2}} = 1.960$

Decision Rule: Reject H_0 if $|z| \geq 1.96$ or Reject H_0 if $p \leq \alpha$

Otherwise, do not reject H_0

An Example: Logistic Regression Inference

4. Compute the test statistic

$$z = \frac{\beta_1}{SE_{\beta_1}} = \frac{0.024}{0.0116} = 2.069$$

5. Conclusion

Reject H_0 since $z \geq 1.96$ or since $p - value \leq \alpha$. We have significant evidence at the $\alpha = 0.05$ level that $\beta_1 \neq 0$. There is evidence of an association between cholesterol level and risk of a coronary event.

The odds ratio for a coronary event is $e^{\beta_1} = 1.02$ for every 1 unit increase in cholesterol. (Or we could say that the odds ratio is 1.27 for every 10-unit increase in cholesterol as this may be a more reasonable and clinically relevant scale to report the results).

We are 95% confident that the true odds ratio is between 1.00 and 1.047. (We could also report the 95% confidence interval for the 10-unit increase instead if we had chosen to present the odds ratio in the previous sentence based on this unit of increase).

R commands: Predict Method for GLM Fits

```
# predicted risk for each patient
risk <- predict(m, type=c("response"))
risk
 1          2          3          4          5          6          7
0.22720323 0.43615030 0.34653363 0.31520904 0.23137263 0.48299515
    0.59393290
        8          9         10         11         12         13
                          14
0.44196119 0.49478598 0.47122323 0.70593666 0.61088437 0.41309669
    0.18133869
...
```

The parameter "type" indicates the type of prediction required.

The default is on the scale of the linear predictors; the alternative "response" is on the scale of the response variable. **Thus for a default binomial model the default predictions are of log-odds (probabilities on logit scale) and type = "response" gives the predicted probabilities.**

R commands: Predict Method for GLM Fits

```
> risk <- predict(m, type=c("response"))


# predicted risk for patient with cholesterol of 190
> risk[41]
 41
0.6808668

# Or manual calculation
> exp(m$coefficients[1]+m$coefficients[2]*190)/(1+exp(m$coefficients[1]+
    m$coefficients[2]*190))
(Intercept)
  0.6808668
```

An Example: Multiple Logistic Regression

Our explanatory variables are cholesterol level, age and gender and our outcome is whether or not the subject had another coronary event.

The p-value for the global test was 0.0058.

A summary of the beta estimates from the model are shown below. Test the global null hypothesis at the $\alpha = 0.05$ level.

Parameter	Estimate	Standard Error	p-value
(Intercept)	-8.536	2.684	0.001
β_{Age}	0.042	0.025	0.096
$\beta_{CholesterolLevel}$	0.029	0.013	0.024
β_{Gender}	2.521	0.803	0.002

An Example: Logistic Regression Inference

The test for the global null hypothesis tests:

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$

$H_1 :$ **at least one** $\beta_i \neq 0$

The p-value for the global test was 0.0058. Since , we reject the null hypothesis and conclude that there is at least one $\beta_i \neq 0$.

Example Regression Coefficient for Age:

$H_0 : \beta_{age} = 0$ **or** $OR_{age} = 1$ (there is no association between age and risk for a coronary event, after controlling for cholesterol level and gender)

$H_1 : \beta_{age} \neq 0$ **or** $OR_{age} \neq 1$ (there is an association between age and risk for a coronary event, after controlling for cholesterol level and gender)

We fail to reject the null hypothesis that or **after adjusting for cholesterol level and gender** since $> \alpha$. We do not have significant evidence at the $\alpha = 0.05$ level that $\beta_{age} \neq 0$ ($p = 0.096$).

The odds ratio for a coronary event is 1.04 for every 1-year increase in age.

An example: logistic regression inference

Cholesterol Level:

Reject $H_0 : \beta_{chol} = 0$ or $OR_{chol} = 1$ after adjusting for age and gender since $p \leq \alpha$

We have significant evidence at the $\alpha = 0.05$ level that $\beta_{chol} \neq 0$.

There is evidence of an association between cholesterol level and risk of a coronary event after adjusting for age and gender.

The odds ratio for a coronary event is 1.029 for every 1-unit increase in cholesterol.

Gender:

Reject $H_0 : \beta_{gender} = 0$ or $OR_{gender} = 1$ after adjusting for age and cholesterol level since $p\alpha$.

There is evidence of an association between gender and risk of a coronary event after adjusting for age and cholesterol level.

The odds ratio for a coronary event is 12.44 for males versus females.

An Example: Multiple Logistic Regression

Use the regression model to **predict the risk of a coronary event for a 60 year old female with a cholesterol level of 150**.

$$
\begin{aligned}
p &= \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}} \\
&= \frac{e^{-8.536+150*0.029+0*2.521+0.042*60}}{1+e^{-8.536+150*0.029+0*2.521+0.042*60}} \\
&= \frac{e^{-1.66}}{1+e^{-1.66}} = 15.9762\%
\end{aligned}
$$

The risk of a coronary event for a 60 year old woman with a cholesterol level of 150 is 15.97%.

Note in the above equation $\boldsymbol{x_{\text{M versus F}} = 0}$ since this is the dummy variable for males (which is equal to 0 for women).

R commands: Generalized linear models (GLMs)

```
> glm(data$event~data$explanatory1 + data$explanatory2 + ... , family=
    binomial)
```

- In multiple logistic regression, use the wald.test() function (from aod package) to get p value for the global test (of all beta coefficients = 0)

Multiple Logistic Regression

```
# multiple logistic regression
> data$male <- ifelse(data$sex =="M", 1, 0)
> m2 <- glm(data$event ~ data$chol + data$male + data$age, family=
    binomial)
> summary(m2)

# overall test
# install.package("aod")

> library(aod)
> wald.test(b=coef(m2), Sigma = vcov(m2), Terms = 2:4)

# Terms: An optional integer vector specifying which coefficients should
     be jointly tested
# Terms defines to compare which regression coefficients,
# here we want to compare the 2 to 4 (first is the intercept)
# It gives as a result Chi-Squared test results, and p-value of it
# if p is smaller than 0.05 you can reject the null hypothesis

# ORs per 1 unit increase
exp(cbind(OR = coef(m2), confint.default(m2)))
```