SUMMER 2025

TEXAS
The University of Texas at Austin

INTRODUCTION TO STATISTICAL MODELING

Center for Biomedical Research Support

LAYLA GUYOT
Assistant Professor of Instruction, Ph.D.
Department of Statistics and Data Sciences
The University of Texas at Austin

Access materials

laylaguyot

Overview | Repositories 8 | Projects

Layla Guyot
laylaguyot

Statistics and Data Science enthusiast: teacher and researcher in education, focusing on bridging the gap between academia and industry.

https://github.com/laylaguyot/CBRS_Intro_Statistical_Modeling

README

Introduction to Statistical Modeling

Materials

Download the slides and corresponding .Rmd files each day

Access RStudio

Go to https://gsafcomp02.ccbb.utexas.edu and select RStudio

Tentative Schedule

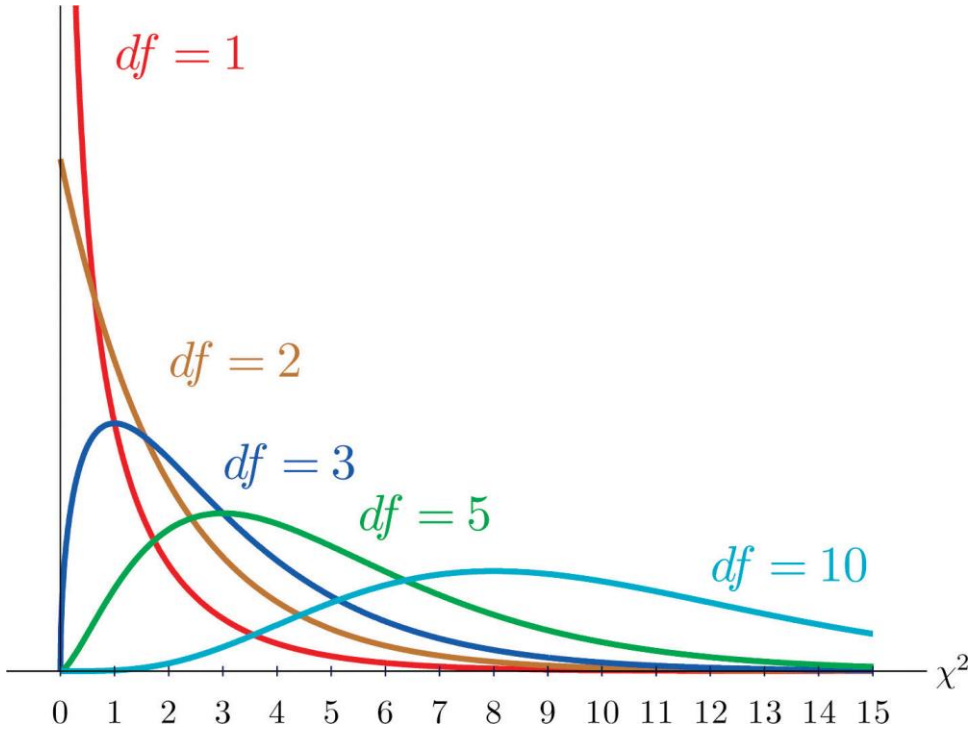
Day 1 **Exploring Data**	• Study design and variables • Descriptive statistics and visualizations • Introduction to hypothesis testing
Day 2 **Making Inferences**	• Probability, random variables, and common probability distributions • Sampling distributions and Central Limit Theorem • Confidence intervals, t-tests, ANOVA, and Chi-square tests
Day 3 **Linear Regression**	• Simple Linear Regression • Multiple Regression with different types of predictors • Model assumptions, evaluation, and comparisons
Day 4 **Logistic Regression**	• Odds • Logistic Regression • Model evaluation with ROC curves or confusion matrix
Day 5 **Model Building**	• Underfitting, overfitting, and cross-validation • Common pitfalls: multicollinearity, transformations • Missing data

χ^2 distribution

A probability distribution that:

- is always positive
- is skewed to the right
- depends on the degree of freedom

df = number of categories – 1



χ^2 Goodness-of-Fit Test

Comparing population counts to hypothesized counts

1. State your hypotheses

H_0: The distribution of the categories **is** [*specify distribution of each category*]

H_A: The distribution of the categories **is not** [*specify distribution of each category*]

χ^2 Goodness-of-Fit Test

Comparing population counts to hypothesized counts

1. State your hypotheses
2. Calculate the test statistic χ^2 (based on sample data)

$$\chi^2 = \sum \frac{(Observed - Expected\ Count)^2}{Expected}$$

What does a large value of χ^2 indicate about the null hypothesis?

with $Expected\ Count = (Expected\ percentage) \cdot n$

χ^2 Goodness-of-Fit Test

Comparing population counts to hypothesized counts

1. State your hypotheses

2. Calculate the test statistic χ^2 (based on sample data)

Species	Adelie	Chinstrap	Gentoo	Total
Observed	146	68	119	333
Expected				333

χ^2 Goodness-of-Fit Test

Comparing population counts to hypothesized counts

1. State your hypotheses
2. Calculate the test statistic $\boldsymbol{\chi^2}$ (based on sample data)
3. Compare test statistic to null distribution (calculate ***p*-value**)
4. Make a conclusion in context, reporting the appropriate statistics (χ^2, *df*, *p*-value).

$$df = number\ of\ categories - 1$$

χ^2 Goodness-of-Fit Test

Comparing population counts to hypothesized counts

Check assumptions:

✓ Random sample

✓ Independent observations

✓ Must have sufficient sample size for:

- All expected counts to be greater than 1
- At least 80% of expected counts are ≥ 5

χ^2 Test of Independence

Comparing population counts for different groups

1. State your hypotheses

H_0: The two variables **are** independent

H_A: The two variables **are not** independent

χ^2 Test of Independence

Comparing population counts for different groups

1. State your hypotheses

2. Calculate the test statistic $\boldsymbol{\chi^2}$ (based on sample data)

$$\boldsymbol{Expected\ Count_{ij} = \frac{(row\ total)_i \cdot (column\ total)_j}{grand\ total}}$$

$$\boldsymbol{\chi^2 = \sum \frac{\left(Observed_{ij} - Expected_{ij}\right)^2}{Expected_{ij}}}$$

χ^2 Test of Independence

Comparing population counts for different groups

1. State your hypotheses
2. Calculate the test statistic $\boldsymbol{\chi^2}$ (based on sample data)
3. Compare test statistic to null distribution (calculate ***p*-value**)
4. Make a conclusion in context, reporting the appropriate statistics (χ^2, *df*, *p*-value).

$$\boldsymbol{df = (number\ of\ categories - 1)(number\ of\ categories - 1)}$$

χ^2 Test of Independence

Comparing population counts for different groups

Check assumptions:

✓ Random sample

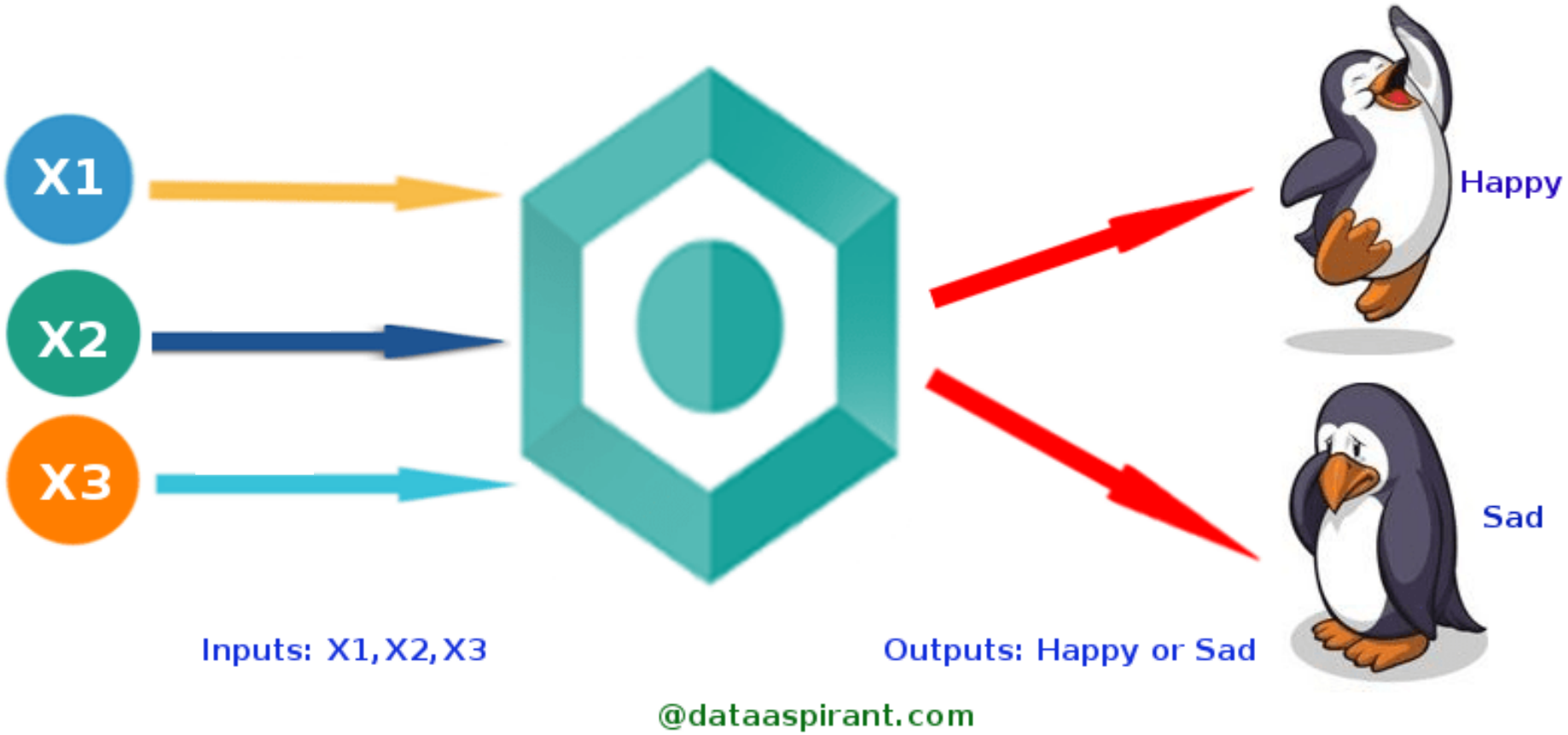
✓ Independent observations

✓ Must have sufficient sample size for:

- All expected counts to be greater than 1
- At least 80% of expected counts are ≥ 5

USING R AND RSTUDIO

Predicting Categorical Outcomes



Introduction to Classification

Confusion Matrix		Predicted	
		Positive	Negative
Outcome	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

$$Accuracy = \frac{True\ Positive\ cases + True\ Negative\ cases}{All\ cases}$$

reports the rate of accurate predicted values

$$True\ Positive\ Rate\ (TPR) = \frac{True\ Positive\ cases}{Positive\ Outcome\ cases}$$

reports the rate of accurate positive predicted values

$$False\ Positive\ Rate\ (FPR) = \frac{False\ Positive\ cases}{Negative\ Outcome\ cases}$$

reports the rate of inaccurate positive predicted values

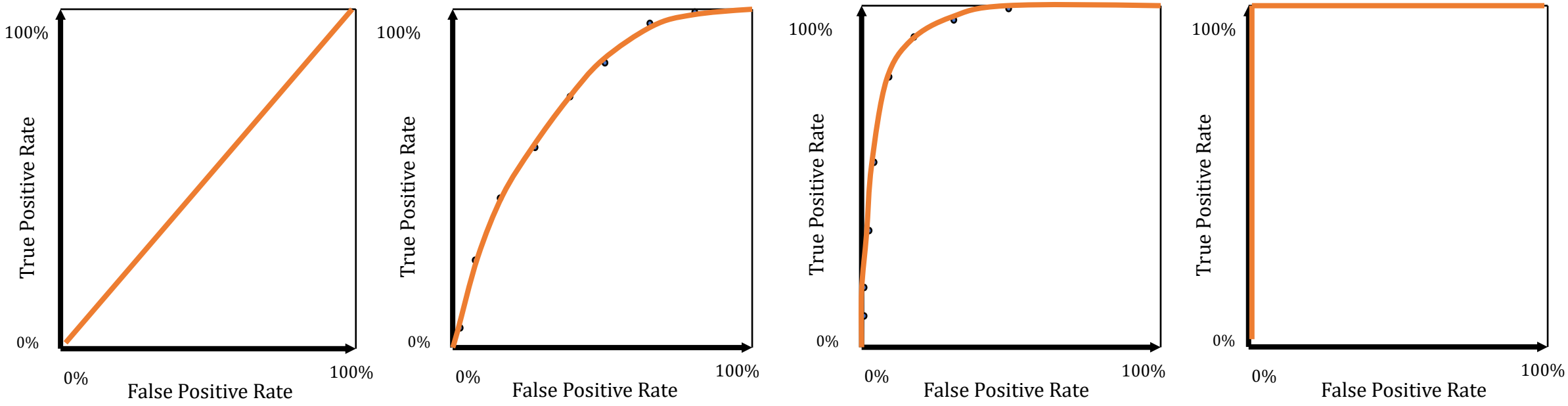
Introduction to Classification

ROC = Receiver Operating Characteristic

- started in electronic signal detection theory (WWII)
- has become popular in biomedical applications (radiology)
- used in machine learning applications to assess classifiers

Introduction to Classification

ROC curve comparison



Random model

Best model

Introduction to Classification

AUC = Area Under the Curve

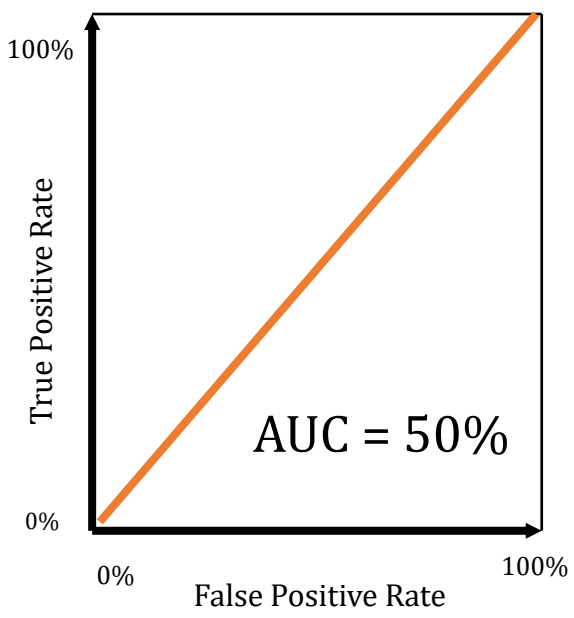
- overall measure of model performance: the higher the area under the curve the better prediction power the model has
- compare different models based on estimated AUC
- interpretation of AUC: a randomly selected individual from the “success” group has a test value larger than for a randomly chosen individual from the “failure” group [AUC] percent of the time.

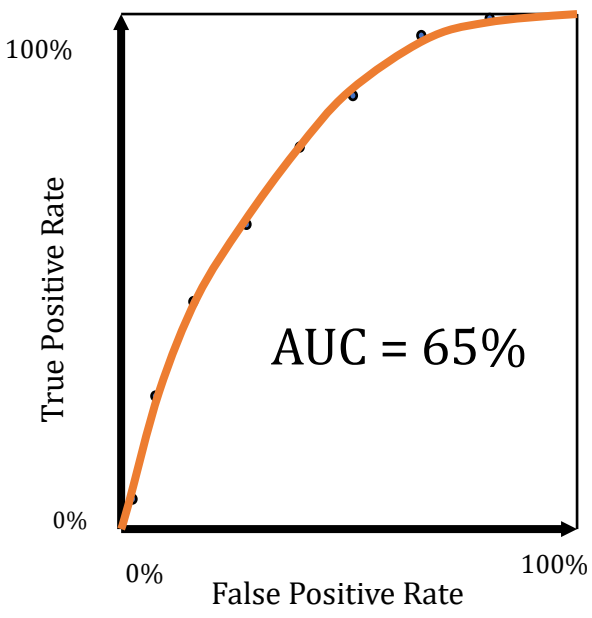
Introduction to Classification

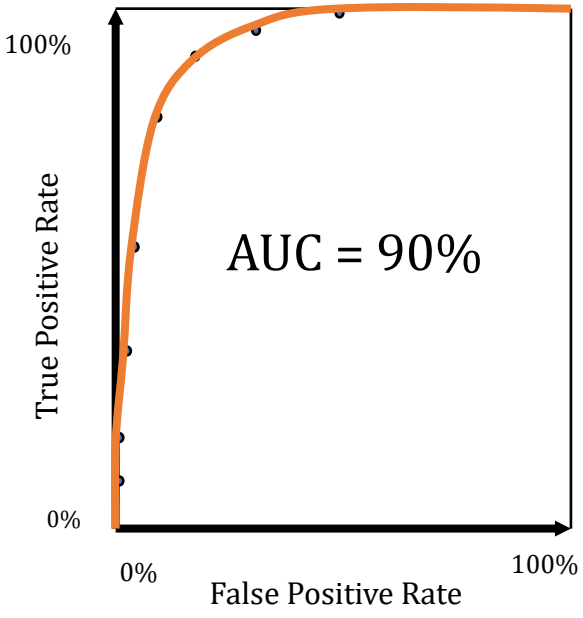
Rules of thumb for AUC

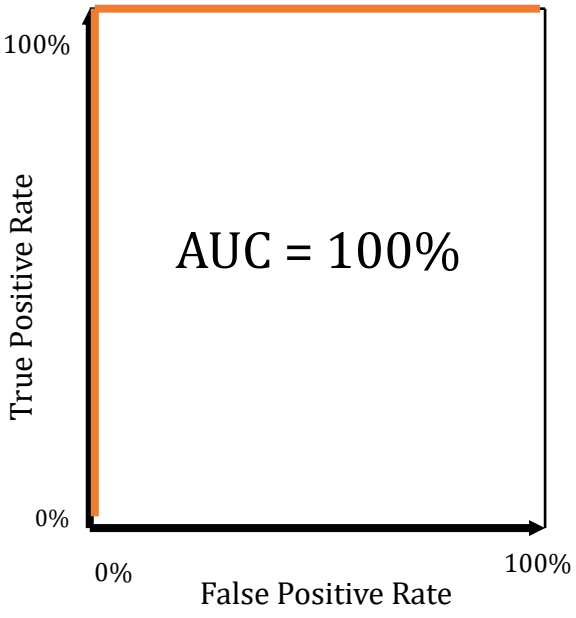
$0.9 - 1.0 : Great!$
$0.8 - 0.9 : Good$
$0.7 - 0.8 : Fair$
$0.6 - 0.7 : Poor$
$0.5 - 0.6 : Bad!$

AUC = Area Under the Curve









Random model

Best model

Logistic Regression

Model

Instead of predicting the mean, we want to predict a probability:

$$\pi = \Pr(Y = 1)$$

Find a *link* function g that takes input values in $[0,1]$ and outputs in $(-\infty, \infty)$:

$$g(p) = \beta_0 + \beta_1 X$$

Odds

Defining odds: $$Odds(success) = \frac{\Pr(success)}{\Pr(no\ success)} = \frac{\pi}{1-\pi}$$

Species	Adelie	Chinstrap	Gentoo	Total
Observed	146	68	119	333

$$Odds(Adelie) = \frac{\frac{146}{333}}{\frac{187}{333}} = 0.78$$

Odds

Defining odds ratio: $$OR = \frac{Odds_1}{Odds_2}$$

	Adelie	Chinstrap	Gentoo	Total
Female	73	34	58	165
Male	73	34	61	168

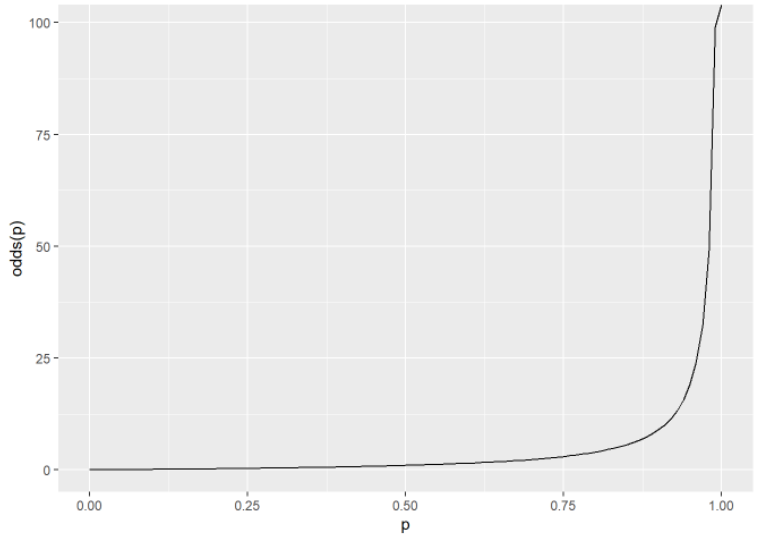
$$OR(being\ male\ in\ Gentoo\ vs\ Adelie) = \frac{\frac{61}{58}}{\frac{73}{73}} = 1.05$$

Logistic Regression

Odds to the logit link function

- Odds of an event: $\text{Odds (success)} = \frac{\Pr(success)}{\Pr(no\ success)} = \frac{\pi}{1-\pi}$

```
##         p   odds
##  [1,] 0.0 0.0000
##  [2,] 0.1 0.1111
##  [3,] 0.2 0.2500
##  [4,] 0.3 0.4286
##  [5,] 0.4 0.6667
```



As p approaches 0, the odds also approach 0

As p approaches 1, the odds go to infinity

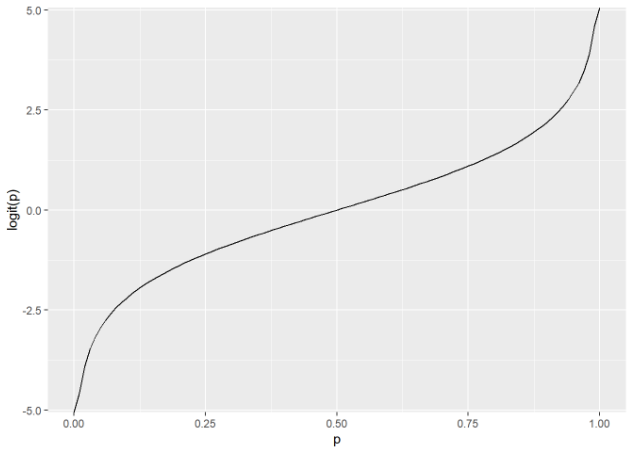
Logistic Regression

Odds to the logit link function

- Odds of an event: $\text{Odds (success)} = \frac{\Pr(success)}{\Pr(no\ success)} = \frac{\pi}{1-\pi}$

- Consider the log of the odds as a function of the probability *p:* $\text{logit}(p) = \ln\left(\frac{\pi}{1-\pi}\right)$

```
##         p   odds   logit
##  [1,] 0.0 0.0000    -Inf
##  [2,] 0.1 0.1111 -2.1972
##  [3,] 0.2 0.2500 -1.3863
##  [4,] 0.3 0.4286 -0.8473
##  [5,] 0.4 0.6667 -0.4055
```



The logit function takes input values in [0,1] and maps into $(-\infty, \infty)$

Logistic Regression

Instead of predicting the mean (linear regression), we predict a probability:

Probability of "success": π_i is the proportion of 1's at any level of X

Two equivalent forms of the logistic regression model:

Logit form

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

Probability form

$$\pi = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

$$\ln(odds) = \beta_0 + \beta_1 X \Rightarrow odds = e^{\beta_0+\beta_1 X}$$

Logistic Regression

Instead of predicting the mean (linear regression), we predict a probability:

Probability of "success": π_i is the proportion of 1's at any level of X

- Logistic regression model: $\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$ — Logit form
- Exponentiate each side: $odds = e^{\beta_0+\beta_1 X} = e^{\beta_0} e^{\beta_1 X}$ — Odds form
- Convert to probabilities: $\pi = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$ — Probability form

Logistic Regression

Interpreting the slope using odds ratio:

What happens when we increase X by 1?

$$e^{\beta_0+\beta_1(X+1)} = e^{\beta_0+\beta_1 X} \cdot e^{\beta_1}$$

When we increase X by 1, the *odds* increase or decrease by a *factor* of e^{β_1} (odds ratio).

Logistic Regression

Example: Let's model the odds of surviving by age with logistic regression.

```
> my_model <- glm(Survival ~ Age, data = titanic, family = binomial(link="logit"))
> summary(my_model)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.136527   0.144715  -0.943    0.345
Age         -0.007899   0.004406  -1.793    0.073 .
```

Logit form

$$\log(\widehat{odds}) = -0.137 - 0.008 \cdot Age$$

Odds form

$$\widehat{odds} = e^{-0.137-0.008 \cdot Age} = e^{-0.137} e^{-0.008 \cdot Age}$$

Logistic Regression

Example: Let's model the odds of surviving by age with logistic regression.

```
> my_model <- glm(Survival ~ Age, data = titanic, family = binomial(link="logit"))
> summary(my_model)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.136527   0.144715  -0.943    0.345
Age         -0.007899   0.004406  -1.793    0.073 .
```

Slope of Age:

Odds form

$$\widehat{odds} = e^{-0.137-0.008 \cdot Age} = e^{-0.137} e^{-0.008 \cdot Age}$$

$$e^{-0.008(1)} \approx 0.992$$

The odds of survival for a passenger are about 99.2% of the odds of survival for a passenger who is one year older.

Logistic Regression

Example: Let's model the odds of surviving by age with logistic regression.

```
> my_model <- glm(Survival ~ Age, data = titanic, family = binomial(link="logit"))
> summary(my_model)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.136527   0.144715  -0.943    0.345
Age         -0.007899   0.004406  -1.793    0.073 .
```

Slope of Age:

Odds form $\widehat{odds} = e^{-0.137-0.008 \cdot Age} = e^{-0.137} e^{-0.008 \cdot Age}$

$e^{-0.008(-1)} \approx 1.008$

easier to interpret odds ratio > 1

The odds of survival of a passenger increases by 1.008 times as age decreases by 1 year.

Logistic Regression

Example: Let's model the odds of surviving by gender with logistic regression.

```
> my_model <- glm(Survival ~ Gender, data = titanic, family = binomial(link="logit"))
> summary(my_model)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9818     0.1040   9.437   <2e-16 ***
GenderMale   -2.4254     0.1360 -17.832   <2e-16 ***
```

Logit form $\log(\widehat{odds}) = 0.98 - 2.43 \cdot Gender$

Odds form $\widehat{odds} = e^{0.98-2.43 \cdot Gender} = e^{0.98} e^{-2.43 \cdot Gender}$

What is the reference group?

Logistic Regression

Example: Let's model the odds of surviving by gender with logistic regression.

```
> my_model <- glm(Survival ~ Gender, data = titanic, family = binomial(link="logit"))
> summary(my_model)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9818     0.1040   9.437   <2e-16 ***
GenderMale   -2.4254     0.1360 -17.832   <2e-16 ***
```

Slope of Gender:

Odds form

$$\widehat{odds} = e^{0.98-2.43 \cdot Gender} = e^{0.98} e^{-2.43 \cdot Gender}$$

$$e^{-2.43(-1)} \approx 11.4$$

easier to interpret odds ratio > 1

Female passengers were about 11.4 times more likely to survive compared to male passengers.

Logistic Regression

Example: Let's model the odds of surviving by gender with logistic regression.

```
> my_model <- glm(Survival ~ Gender, data = titanic, family = binomial(link="logit"))
> summary(my_model)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9818     0.1040   9.437   <2e-16 ***
GenderMale   -2.4254     0.1360 -17.832   <2e-16 ***
```

Predict the odds and probability of surviving for a female passenger.

$$\widehat{odds} = e^{0.98-2.43(0)} = e^{0.98} \approx 2.66$$

$$\hat{p} = \frac{e^{0.98-2.43(0)}}{1+e^{0.98-2.43(0)}} \approx 0.727$$

A female passengers was about 2.66 more likely to survive (than not).
The probability for a female passenger to survive was expected to be about 72.7%.

Logistic Regression

Assumptions:

✓ The log odds have a **linear** relationship with the predictor

✓ The observations are **independent**

✓ Data was collected **randomly**

x **Normality** does not apply because the responses are 0 or 1

x There is no **equal variance** because the variability in Y is greater when π is near ½ and lower when π is near 0 or 1

Logistic Regression

Comparing population slope to 0

1. State your hypotheses

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

Logistic Regression

Comparing population slope to 0

1. State your hypotheses

2. Calculate the test statistic z (based on sample data)

$$z = \frac{b_1}{SE_{b_1}}$$

From a Wald test

Logistic Regression

Comparing population slope to 0

1. State your hypotheses
2. Calculate the test statistic ***z*** (based on sample data)
3. Compare test statistic to null distribution (calculate ***p*-value**)
4. Make a conclusion in context, reporting the appropriate statistics (z, p-value).

Logistic Regression

When reporting results of a significant test, we should report a measure of the effect size with a **confidence interval** of the **population slope:**

$$\boldsymbol{b_1 \pm z^* \cdot SE_{b_1}}$$

Easier to interpret in the context of the *odds*
(increase or decrease by a *factor* of e^{β_1}).

Logistic Regression

Model estimation:

Parameters are chosen to maximize the likelihood of the observed sample (Maximum Likelihood Estimation):

$$L = \prod \hat{\pi}_i^{Y_i}(1-\hat{\pi}_i)^{1-Y_i} \qquad \text{with} \quad \pi_i = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}$$

Maximize L in terms of β_0 and β_1

➔ Maximize $\log L$

Logistic Regression

Model estimation: $L = \prod \hat{\pi}_i^{Y_i}(1-\hat{\pi}_i)^{1-Y_i}$ with $\pi_i = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}$

- Deviance $D = -2\log L$

  We use the deviance as a measure of fit (positive and smaller is better, equivalent to *SSE* for linear regression)

- AIC $AIC = -2\ log\ L + 2p$

  We use the Akaike Information Criterion for comparing models (taking into account the number of predictors).

- Pseudo R^2 (McFadden)

$$R^2 = 1 - \frac{-2\log L}{-2\log L_0} = 1 - \frac{\log L}{\log L_0}$$

We use the Pseudo R^2 as a measure of the **improvement in model fit** relative to the null model (the closer to 1 the better).

Logistic Regression

Assumptions:

✓ The log odds have a **linear** relationship with the predictor

✓ Data was collected **randomly**

✓ The observations are **independent**

Compare to linear regression models:

x **Normality** does not apply because the responses are 0 or 1

x There is no **equal variance** because the variability in Y is greater when p is near ½ and lower when p is near 0 or 1

BREAK TIME

BACK AT ...

More on Logistic Regression

Other logistic regression models

- With multiple predictors $\ln(odds) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
  - ➢ Slope interpretation: Controlling for / Holding other predictors constant
- With interactions $\ln(odds) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 * X_2$
  - ➢ Slope interpretation: Set the other predictor in the interaction to be zero

More on Logistic Regression

Example: How was surviving the sinking of the Titanic affected by age, gender and class?

```
Survival ~ Age + Gender + passengerClass
```

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.522072   0.326701  10.781  < 2e-16 ***
Age               -0.034393   0.006331  -5.433 5.56e-08 ***
GenderMale        -2.497845   0.166037 -15.044  < 2e-16 ***
passengerClass2nd -1.280567   0.225538  -5.678 1.36e-08 ***
passengerClass3rd -2.289658   0.225802 -10.140  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1414.62  on 1045  degrees of freedom
Residual deviance:  982.45  on 1041  degrees of freedom
  (263 observations deleted due to missingness)
AIC: 992.45
```

Significance

While controlling for other predictors present in the model, each predictor contributes significantly to predicting the survival of a passenger.

More on Logistic Regression

Example: How was surviving the sinking of the Titanic affected by age, gender and class?

```
Survival ~ Age + Gender + passengerClass
```

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.522072   0.326701  10.781  < 2e-16 ***
Age              -0.034393   0.006331  -5.433 5.56e-08 ***
GenderMale       -2.497845   0.166037 -15.044  < 2e-16 ***
passengerClass2nd -1.280567   0.225538  -5.678 1.36e-08 ***
passengerClass3rd -2.289658   0.225802 -10.140  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1414.62  on 1045  degrees of freedom
Residual deviance:  982.45  on 1041  degrees of freedom
  (263 observations deleted due to missingness)
AIC: 992.45
```

Interpret slopes

The odds of surviving decrease for male passengers compared to female passengers, while controlling for age and class.

```
> exp(my_model$coefficients)
      (Intercept)                Age         GenderMale passengerClass2nd passengerClass3rd
      33.85451255         0.96619149         0.08226204         0.27787957         0.10130106
```

More on Logistic Regression

Example: How was surviving the sinking of the Titanic affected by age, gender and class?

```
Survival ~ Age + Gender + passengerClass
```

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.522072   0.326701  10.781  < 2e-16 ***
Age              -0.034393   0.006331  -5.433 5.56e-08 ***
GenderMale       -2.497845   0.166037 -15.044  < 2e-16 ***
passengerClass2nd -1.280567   0.225538  -5.678 1.36e-08 ***
passengerClass3rd -2.289658   0.225802 -10.140  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1414.62  on 1045  degrees of freedom
Residual deviance:  982.45  on 1041  degrees of freedom
  (263 observations deleted due to missingness)
AIC: 992.45
```

Interpret slopes

The odds of surviving decrease for passengers in 2nd class, and even more for passengers in 3rd class, compared to passengers in 1st class, while controlling for age and gender.

```
> exp(my_model$coefficients)
      (Intercept)                Age         GenderMale passengerClass2nd passengerClass3rd
      33.85451255         0.96619149         0.08226204         0.27787957         0.10130106
```

USING R AND RSTUDIO

Next

Day 5
Model Building

- Underfitting, overfitting, and cross-validation
- Common pitfalls: multicollinearity, transformations
- Missing data

Any questions? comments?