

# C4C: Model Evaluation and Comparison

Tatum Katz and Sam Sambado  
08-22-2022

# Overview

1. **Why should I evaluate my models?**

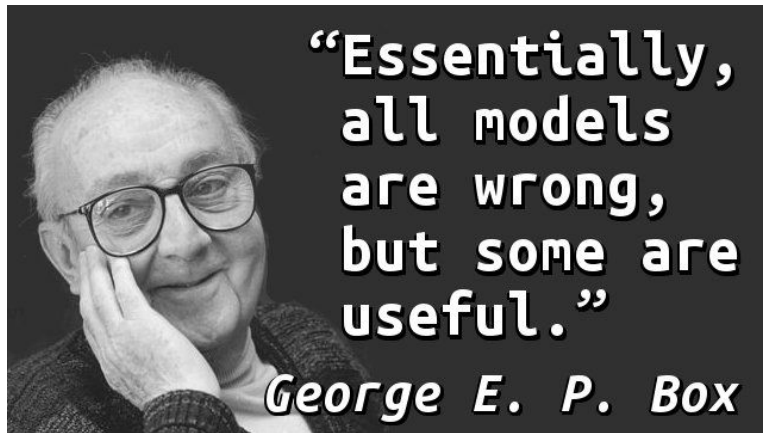
2. **Toolbox:**

- a. Power analysis (anything)
- b. Residual analysis (LM and GLM)
- c. AIC, BIC (any models built on same data)
- d. Confusion matrix (binary outcome variable)
- e. Cross validation, k-fold validation, out-of-bag/in-bag (just about anything)



3. **Discussion:** how will you evaluate your model(s)?

# Why should I evaluate my models?

- Models are approximations of reality
- In a frequentist framework, the phenomenon can be perfectly described by a single model
- We can only estimate that perfect model
- We need to know how good/bad of a job we are doing!



# Goodness of fit vs. complexity

- **Goodness of fit** and **complexity** are two concepts we use to compare models
  -  **Goodness of fit** describes how well the model fits the data (i.e., how small are the residuals?)
  -  **Complexity** describes how many parameters are in the model → *OCCAM'S RAZOR anything not necessary should not be included*
- We want models with high goodness of fit and low complexity
- This is hard
  - Its a trade-off
  - Bias vs. variance
- Tests/metrics used to compare models must take into account both

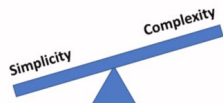
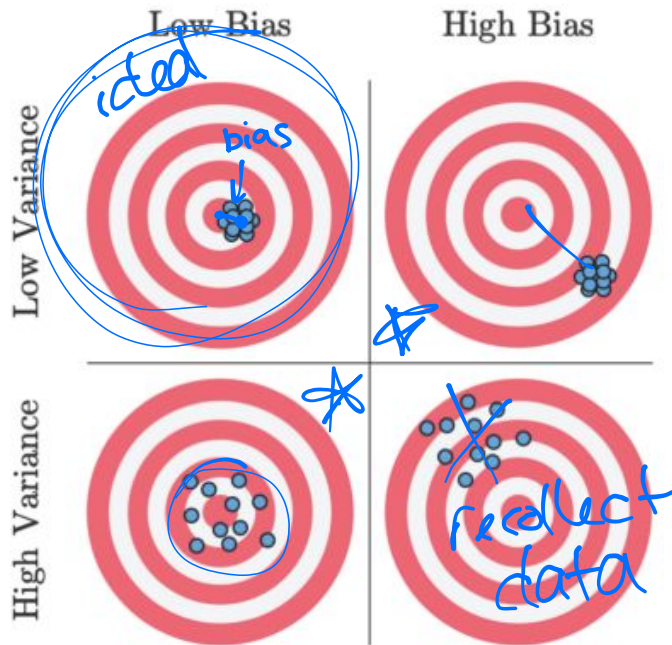


# Goodness of fit vs. complexity

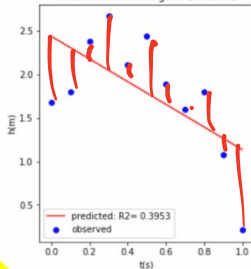
↑GoF  
↑comp.



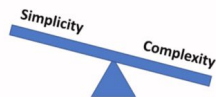
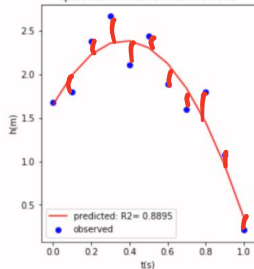
A custom bed.  
Great for you! Horrible for others



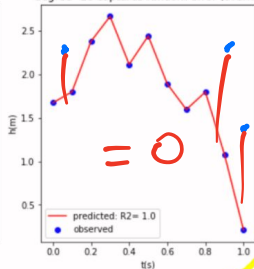
linear model is not good (underfit)



quadratic model is what we need



degree=10 captures random error (overfit)



# Inherent bias and model validation

- **We all have internal biases which may affect our model**
  - What variables you choose to include/exclude
  - What likelihood you select based on previous experience
  - Selecting a model type with a low computational cost

# Inherent bias and model validation

- **We all have internal biases which may affect our model**
  - What variables you choose to include/exclude
  - What likelihood you select based on previous experience
  - Selecting a model type with a low computational cost
- **Every single piece of your model should be validated using model comparison!**
  - Parameters
  - Likelihoods
  - Tuning parameters
  - ANYTHING!

# Inherent bias and model validation

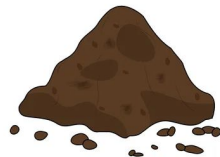
- **We all have internal biases which may affect our model**
  - What variables you choose to include/exclude
  - What likelihood you select based on previous experience
  - Selecting a model type with a low computational cost
- **Every single piece of your model should be validated using model comparison!**
  - Parameters
  - Likelihoods
  - Tuning parameters
  - ANYTHING!
- Every question can be turned into a hypothesis test by comparing to a **null** or **full** model

# Null and full models

**Example:** I want to know what environmental parameters predict the proportion of red vs. blue flowers in a field

I measure:

- Rain
- Temperature
- Wind
- Soil composition



# Null and full models

I hypothesize that rain and temperature are the most important predictors, but I need to validate my choice.

I can create a **null** and a **full** model to compare against!

*includes all variables*

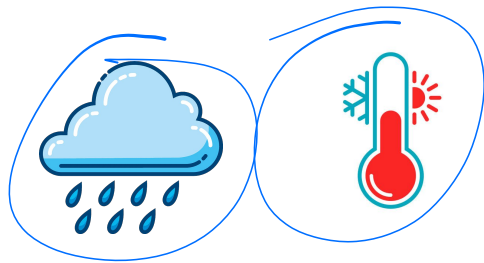
**Full model:** proportion red flowers  $\sim$  rain + temperature + wind + soil

*no variables*

**Null model:** proportion red flowers  $\sim$  1

*only estimate the mean*


*♥* **My model:** proportion red flowers  $\sim$  rain + temperature



# Null and full models

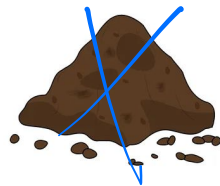
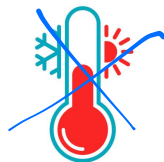
**Full model:** proportion red flowers  $\sim$  rain + temperature + wind + soil

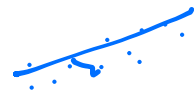
**Null model:** proportion red flowers  $\sim 1$  

 **My model:** proportion red flowers  $\sim$  rain + temperature

I can calculate a statistic or metric for each model, and see which gets the best score.

This would provide evidence for which model is the best.



linear  
w/ 

Poisson  
neg bin  
gamma

# Model Evaluation and Comparison Toolbox

- Lots of ways to validate a model
- Some methods only work for certain models
  - Not every tool works for problem!



# Model Evaluation and Comparison Toolbox

- Lots of ways to validate a model
- Some methods only work for certain models
  - Not every tool works for problem!
- My rule of thumb: use all appropriate methods, compare and see if they all pick the same model
  - If they don't, think about why - can reveal new insights



# Model Evaluation and Comparison Toolbox

- Lots of ways to validate a model
- Some methods only work for certain models
  - Not every tool works for problem!
- My rule of thumb: use all appropriate methods, compare and see if they all pick the same model
  - If they don't, think about why - can reveal new insights
- My goal today: show you all the tools, when to use them, and how to interpret them



# Toolbox: Power Analysis

**Statistical power:** the probability that the test correctly rejects the null hypothesis

- **Low statistical power:** large risk of Type II error i.e. false negative
- **High statistical power:** small risk of Type II error



*control @ 0.05*

*$1 - \beta = \text{power}$   
power  $\sim 80\%$*

		Reality	
		Positive	Negative
Study Finding	Positive	<b>True Positive</b> (Power) ( $1 - \beta$ )	False Positive <b>Type I Error</b> ( $\alpha$ )
	Negative	False Negative <b>Type II Error</b> ( $\beta$ )	<b>True Negative</b>

# Toolbox: Power Analysis



**Statistical power:** the probability that the test correctly rejects the null hypothesis

- **Low statistical power:** large risk of Type II error i.e. false negative
- **High statistical power:** small risk of Type II error

Power is determined by the relationship between:

- **Effect size** (magnitude of difference between the means)
- **Variability** (how much variance there is within each sample and between samples)

		Reality	
		Positive	Negative
Study Finding	Positive	<b>True Positive</b> (Power) ( $1-\beta$ )	False Positive <b>Type I Error</b> ( $\alpha$ )
	Negative	False Negative <b>Type II Error</b> ( $\beta$ )	<b>True Negative</b>

# Toolbox: Power Analysis



Power analysis tells us how to get the level of power we desire, or what our power actually is

- Can be done before the experiment to determine sample size
  - For a given difference in means, for a given value of standard deviation
  - But, we need to know the variability, which can be hard
- Can be done after the experiment (post-hoc) to determine the power of the test
  - Often done if the result was not significant, to make sure this wasn't just a result of small sample size
  - If the null result is “desired”, can be used to show that null results are valid and not just due to sample size

Useful for model evaluation! How do you know that your results aren't just because your sample size was too small?

# Toolbox: Power Analysis

for the model,  
given the  
data



likelihood

Power has four related parts:

1. Effect size: the magnitude of the result in the population
2. Sample size: how many observations in the sample,  $n$
3. Significance level  $\alpha$ : the level at which you will run the test (type 1 error prob)  
 $0.05$
4. Statistical power  $\beta$ : the probability of correctly rejecting a false null  
 $0.80$

For every statistical test and model, there exists a formula to calculate any one of these four values given the other three.

power . t . test (data)

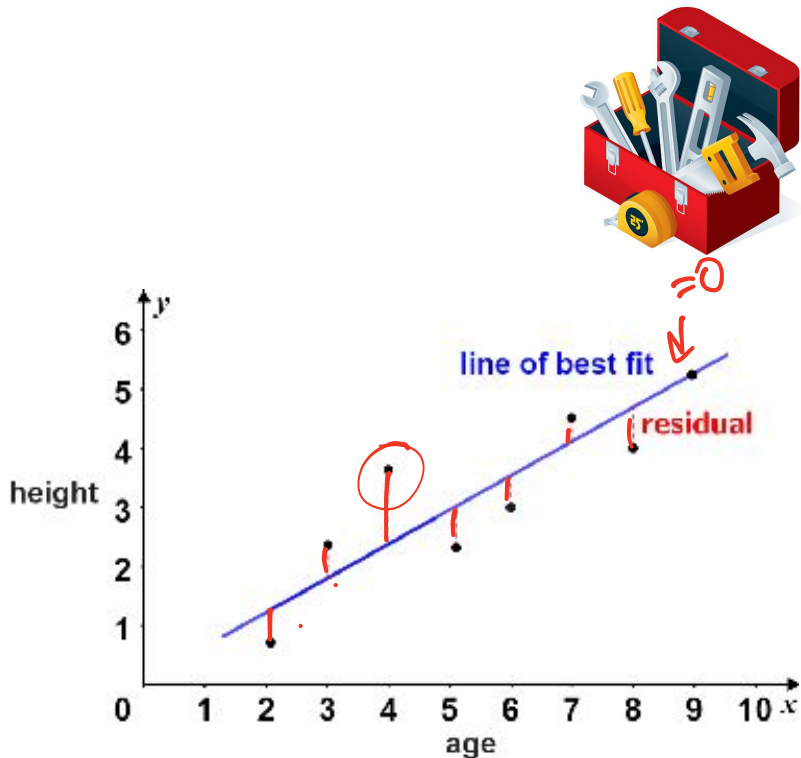
# Toolbox: Residual Analysis

**Residuals** are the difference between the model estimates and the actual data

*error in the model*  
↳ measurement error  
↳ model error

They can tell you about the bias and variance (goodness of fit and complexity) of your model

Residuals for a linear model are simple;  
residuals for a GLM get more tricky

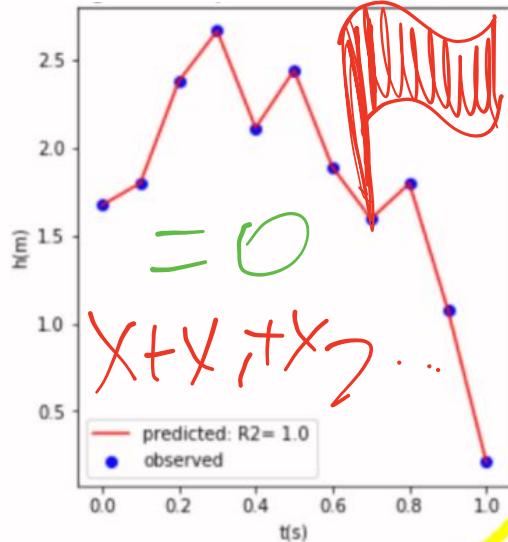
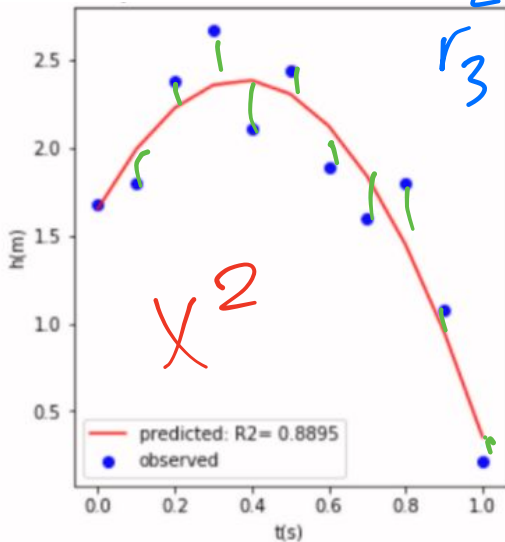
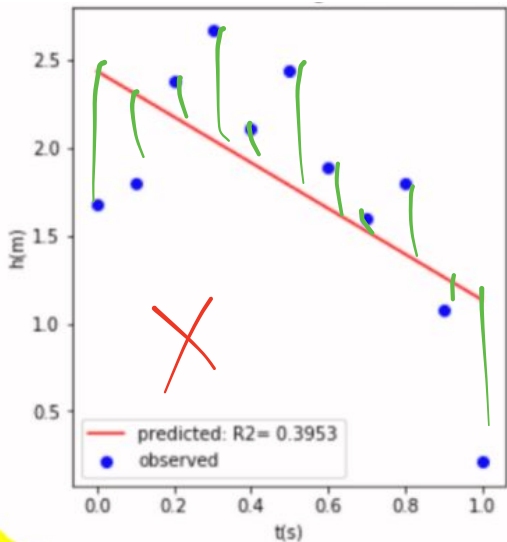


# Toolbox: Residual Analysis

residuals (m. object)



$r_1$   
 $r_2$   
 $r_3$



# Toolbox: Residual Analysis



Visualizing linear model residuals is a good way to check that you are meeting the assumptions of your model

In R: `plot(lm.object)`

Assumptions of a linear model:

1. Random, independent data (can't check)
2. Equal variances
3. Normal distribution of the residuals

# Toolbox: Residual Analysis

Assumptions of a linear model:

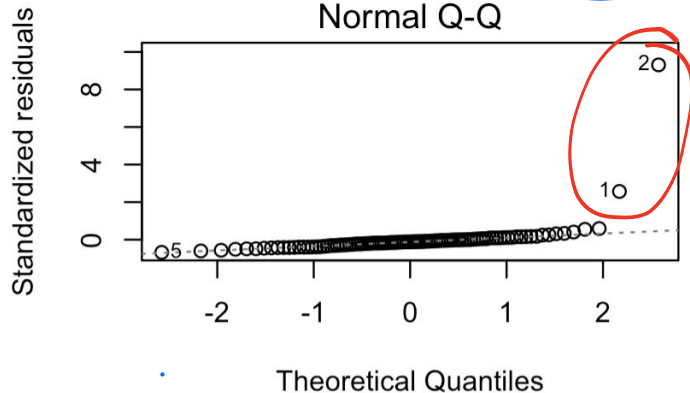
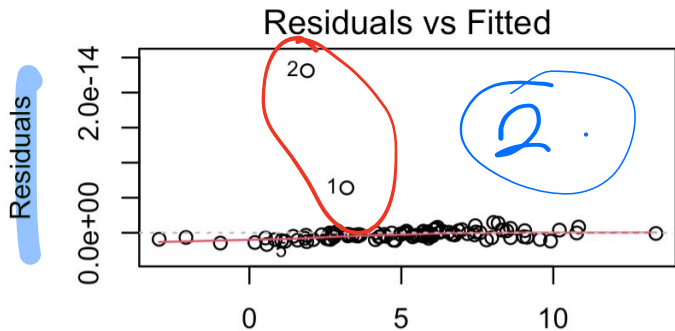
1. Random, independent data (can't check)
2. Equal variances ✓
3. Normal distribution of the residuals ✓

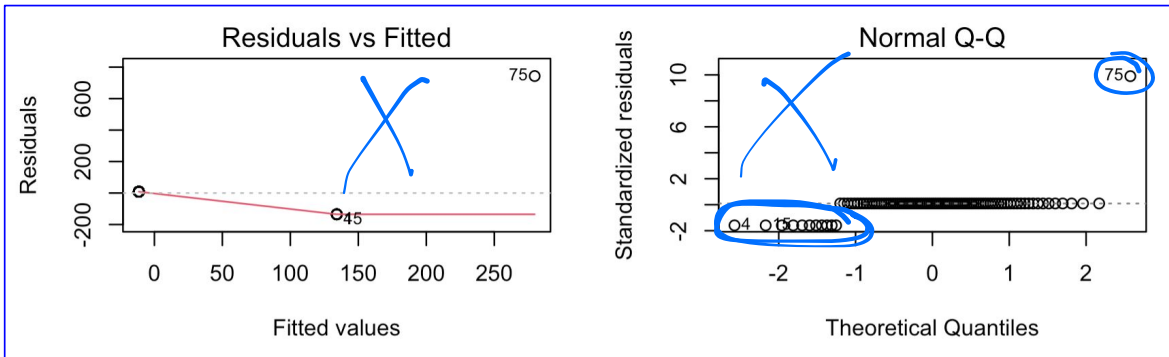
plot()

Q-Q plot



3.





# Toolbox: Residual Analysis



Summarizing linear model residuals:

- $R^2$ : the proportion of variance in the data as explained by the model  
 $R^2 \geq 0.5$ 
  - “My model has an  $R^2$  of .35, therefore, my model explains 35% of the variance in my data”
  - ~~Adjusted  $R^2$ : no direct interpretation, takes into account number of parameters, suggested not to use~~
- **F ratio**: a statistical test to see if one model explains more variance than the other

Both are given in `summary(lm.object)` output in R!

$$R^2 = 1 - \frac{RSS}{TSS}$$

$R^2$  = coefficient of determination

—  $RSS$  = sum of squares of residuals

—  $TSS$  = total sum of squares

$$F = \frac{\frac{\text{Sum Squares Model}}{\text{df Model}}}{\frac{\text{Sum Squares Error}}{\text{df Error}}}$$

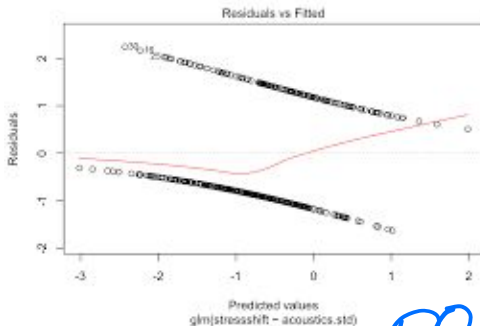
# Toolbox: Residual Analysis

## GLM Residuals

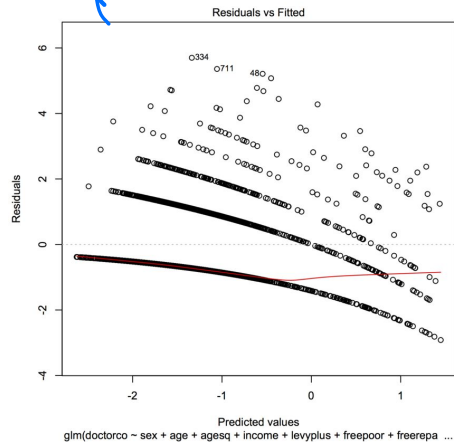
- For each distribution, residuals will look different
- Best plan of action:
  - simulate fake data from the model (use `predict()` in R) and visualize residuals
  - If your model fits well, your real residuals should look like the simulated ones
- GLM variations of  $R^2$  exist for some distributions, but not all

neg. binomial

binomial



Poisson



↻  
Toolbox: AIC, BIC

tatum's favorite



## Akaike's Information Criterion and Bayes' Information Criterion

- Works for any models, as long as they are built using the same data
- Build as many models (on the same data) as you like, and compare to see who has the lowest AIC and BIC!
- Based on the likelihood: given the model, how likely is it that the data came from it?
  - i.e., if your model was the “truth”, how well do the data fit it?

# Toolbox: AIC, BIC



- **AIC:** Akaike's Information Criterion

- Gives a score for how good the model fits the data, penalized by how many parameters are in the models
- The lower the score, the better
- 

- **BIC:** Bayes' Information Criterion

- Same as AIC but with a larger penalty for more parameters

- more conservative  
- pick simpler models

$$AIC = 2k - 2\ln(\hat{L}) \quad BIC = k\ln(n) - 2\ln(\hat{L})$$

# params      likelihood      sample size

AIC  
BIC

# Toolbox: Confusion Matrix



That is a **confusion matrix** ->

- AKA **error matrix**
- Type of **contingency table**
- “Agreement between two raters”

		Reality	
		Positive	Negative
Study Finding	Positive	<b>True Positive</b> (Power) (1- $\beta$ )	False Positive <b>Type I Error</b> ( $\alpha$ )
	Negative	False Negative <b>Type II Error</b> ( $\beta$ )	<b>True Negative</b>

# Toolbox: Confusion Matrix



That is a **confusion matrix** ->

- Useful for classification models
  - logistic regression
  - multinomial outcome regression
  - Many types of machine learning

		Predicted condition	
		Cancer	Non-cancer
Actual condition	Total	7	5
	8 + 4 = 12		
	Cancer	6	2
	8		
	Non-cancer	1	3
	4		

# Toolbox: Confusion Matrix

Can calculate many informative metrics:

- Accuracy, sensitivity, specificity...



		Predicted condition			
		Positive (PP)	Negative (PN)	Informedness, bookmaker informedness (BM) $= \text{TPR} + \text{TNR} - 1$	Prevalence threshold (PT) $= \frac{\sqrt{\text{TPR} \times \text{FPR}} - \text{FPR}}{\text{TPR} - \text{FPR}}$
Actual condition	Total population $= P + N$				
	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{\text{TP}}{P} = 1 - \text{FNR}$	False negative rate (FNR), miss rate $= \frac{\text{FN}}{P} = 1 - \text{TPR}$
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection	False positive rate (FPR), probability of false alarm, fall-out $= \frac{\text{FP}}{N} = 1 - \text{TNR}$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{\text{TN}}{N} = 1 - \text{FPR}$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{\text{TP}}{\text{PP}} = 1 - \text{FDR}$	False omission rate (FOR) $= \frac{\text{FN}}{\text{PN}} = 1 - \text{NPV}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Negative likelihood ratio (LR-) $= \frac{\text{FNR}}{\text{TNR}}$
	Accuracy (ACC) $= \frac{\text{TP} + \text{TN}}{P + N}$	False discovery rate (FDR) $= \frac{\text{FP}}{\text{PP}} = 1 - \text{PPV}$	Negative predictive value (NPV) $= \frac{\text{TN}}{\text{PN}} = 1 - \text{FOR}$	Markedness (MK), deltaP ( $\Delta p$ ) $= \text{PPV} + \text{NPV} - 1$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR-}}$
	Balanced accuracy (BA) $= \frac{\text{TPR} + \text{TNR}}{2}$	F <sub>1</sub> score $= \frac{2\text{PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$	Fowlkes–Mallows index (FM) $= \sqrt{\text{PPV} \times \text{TPR}}$	Matthews correlation coefficient (MCC) $= \sqrt{\text{TPR} \times \text{TNR} \times \text{PPV} \times \text{NPV}} - \sqrt{\text{FNR} \times \text{FPR} \times \text{FOR} \times \text{FDR}}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}$

Sources: [21][22][23][24][25][26][27][28][29] [view](#) · [talk](#) · [edit](#)

# Toolbox: Confusion Matrix



Some metrics you can calculate:

- **Accuracy** (ACC) =  $\frac{TP + TN}{P + N}$
- **Sensitivity** (SEN, true positive rate, power) =  $\frac{TP}{P}$
- **Specificity** (SPC, true negative rate) =  $\frac{TN}{N}$

1 is perfect

		Predicted condition	
		Positive (PP)	Negative (PN)
Actual condition	Total population = P + N		
	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection

# Toolbox: Confusion Matrix



## Cohen's Kappa

- Ranges -1 to 1
    - 0 is random chance, 1 is perfect agreement, -1 is perfect disagreement
  - A nice way to summarize the confusion matrix
  - In R, you can generate a confusion matrix and then use a single command to calculate and interpret kappa (including a p-value!)
- 0 = agreement equivalent to chance.
  - 0.1 – 0.20 = slight agreement.
  - 0.21 – 0.40 = fair agreement.
  - 0.41 – 0.60 = moderate agreement.
  - 0.61 – 0.80 = substantial agreement.
  - 0.81 – 0.99 = near perfect agreement
  - 1 = perfect agreement.

# Toolbox: Confusion Matrix



BUT WAIT! Logistic regression doesn't give 0s and 1s as predicted values... it gives probabilities

- What probability should count as a 1 or a 0?
- What is the **threshold**?

# Toolbox: Confusion Matrix



BUT WAIT! Logistic regression doesn't give 0s and 1s as predicted values... it gives probabilities

- What probability should count as a 1 or a 0?
- What is the **threshold**?

0.5

Calculate the metrics over a range of thresholds and compare the results.

`optimal.thresholds()` can help

0.12 , 0.89

Often, a threshold which maximizes accuracy, sensitivity, or specificity is chosen.

# Toolbox: Confusion Matrix



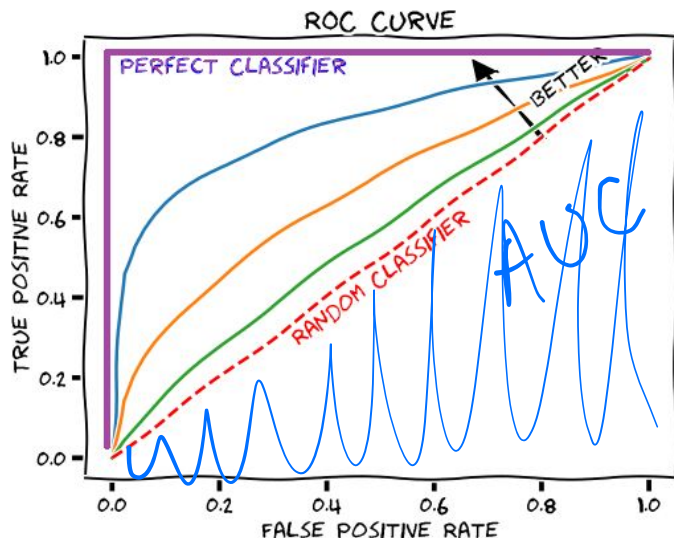
What if I don't want to pick a threshold?

**Area Under the Curve of the Receiver Operator Curve (AUC of the ROC)**

- 1 = perfect classifier
- 0.5 = random classifier

$$AUC \geq 0.70$$

Package **pROC** is great for this



# Toolbox: Cross-validation



**Cross-validation** is a tool to validate your model using another set of data

- **Train data:** the data you use to build your model (in-bag)
- **Test data:** data you test the performance of your model with (out-of-bag)

If your model is good, it should be able to accurately predict new data

If the predicted data is very similar to the test data, then you probably have a good model

# Toolbox: Cross-validation



Two types of test data:

- Data that were withheld during initial analysis
  - Usually save ~20% of dataset for test data
  - But if your training data have biases, so your model has biases, you won't detect them

# Toolbox: Cross-validation



Two types of test data:

- Data that were withheld during initial analysis
  - Usually save ~20% of dataset for test data
  - But if your training data have biases, so your model has biases, you won't detect them
- Data that were independently collected
  - Very hard to find
  - Can help find biases in training data

# Toolbox: Cross-validation

*train + test*

How to:

1. Get predictions for your model (use function `predict()`)
2. Analyze:
  - a. Which data points does the model predict well?
  - b. Which data are poorly predicted?



# Toolbox: Cross-validation



**K-fold cross-validation** is useful when you don't have independent data

*1 data point per fold!*

- Training data is split into K folds (blocks) and each fold is use to train and test the model
- Can calculate a model evaluation metric for each of K models and average them to better understand your model
- Can see which models perform better or worse

# Toolbox: Cross-validation



**K-fold cross-validation** is useful when you don't have independent data

How to:

1. Select a metric to validate your model (**not** AIC or BIC!)
2. Split data into K blocks (folds), as few as one point per block
3. Train the model on each fold of data, and calculate K metrics for K models

There are some R packages that will do this for you - depends on your model type

# Discussion

Take 3 minutes to write down:

1. Questions you have on today's lecture
2. What tools do you think you can use to evaluate your model?
3. Do you foresee any model evaluation issues?

Then, we will discuss!