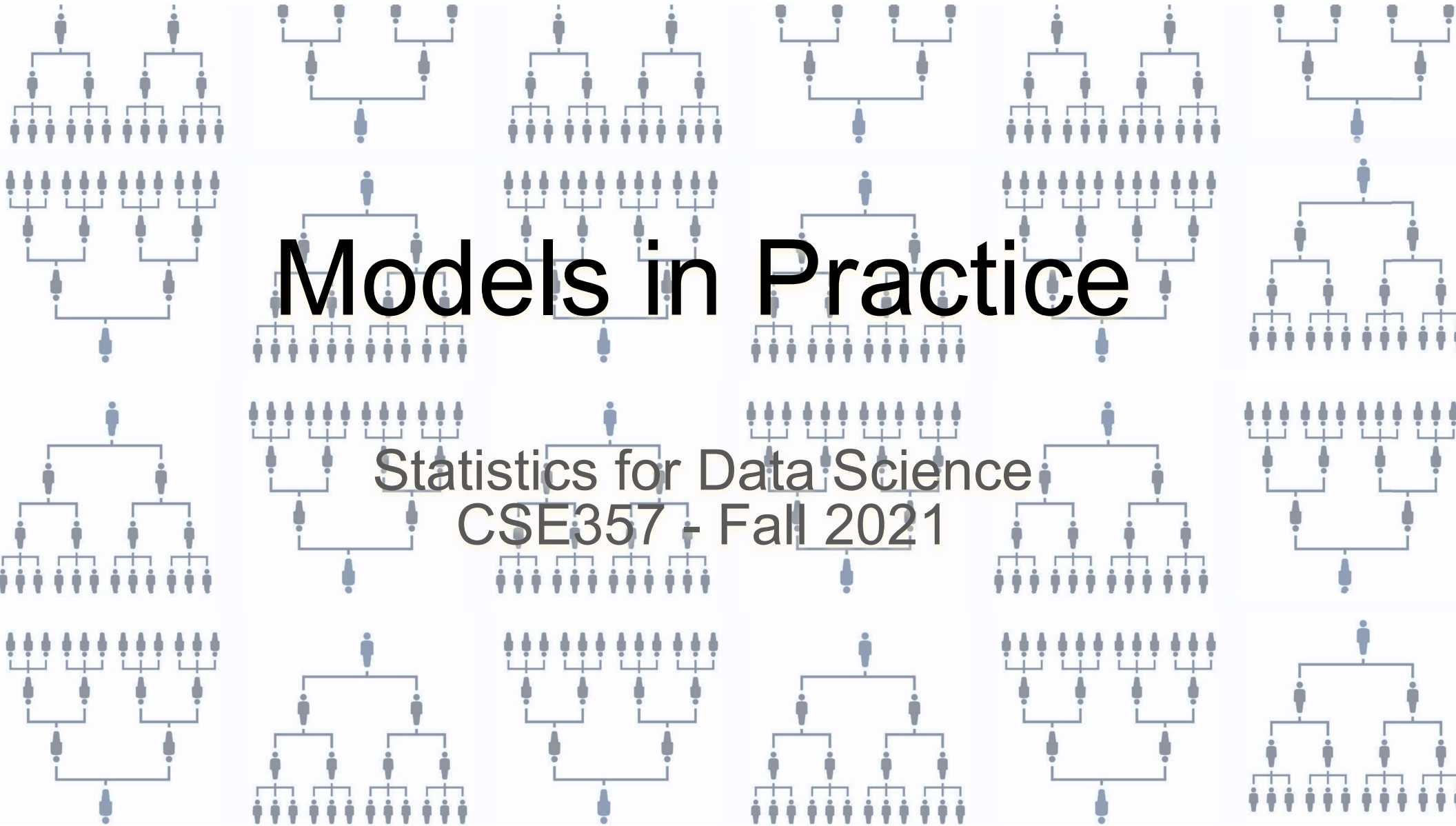


Models in Practice

Statistics for Data Science
CSE357 - Fall 2021

Statistical Models in Practice

Accuracy Metrics

Moderation Analysis

Mediation Analysis

Multi-Level Analysis (Fixed Effect Model)

Accuracy Metrics

(Wikimedia, 2021)

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	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

Accuracy Metrics

(Wikimedia, 2021)

Sources: [3][4][5][6][7][8][9][10] view · talk · edit

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**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
		Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	
		False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	
		F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	

Accuracy Metrics

(Wikimedia, 2021)

Sources: [3][4][5][6][7][8][9][10] view · talk · edit

		Predicted condition			
	Total population = P + N	**Positive (PP)**	**Negative (PN)**		
Actual condition	**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 $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP)**, type I error, false alarm, overestimation	**True negative (TN)**, correct rejection		
		Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$			
		False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$			
		F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$			

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

(harmonic mean of precision and recall)

Accuracy Metrics

(Wikimedia, 2021)

Sources: [3][4][5][6][7][8][9][10] view · talk · edit

		Predicted condition			
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**		
Actual condition	**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 $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP),** type I error, false alarm, overestimation	**True negative (TN),** correct rejection		
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$			
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$			
	Balanced accuracy $(BA) = \frac{TPR + TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$			

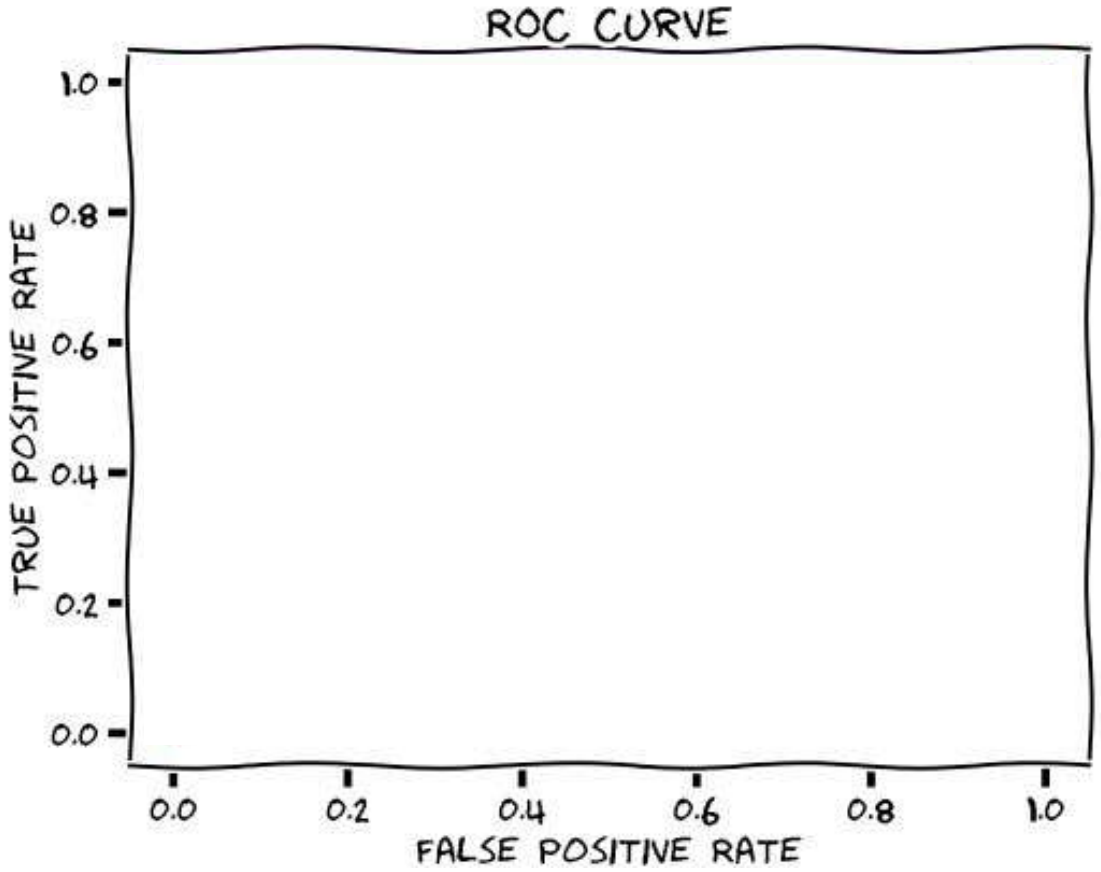
Accuracy Metrics

(Wikimedia, 2021)

Sources: [3][4][5][6][7][8][9][10] view · talk · edit

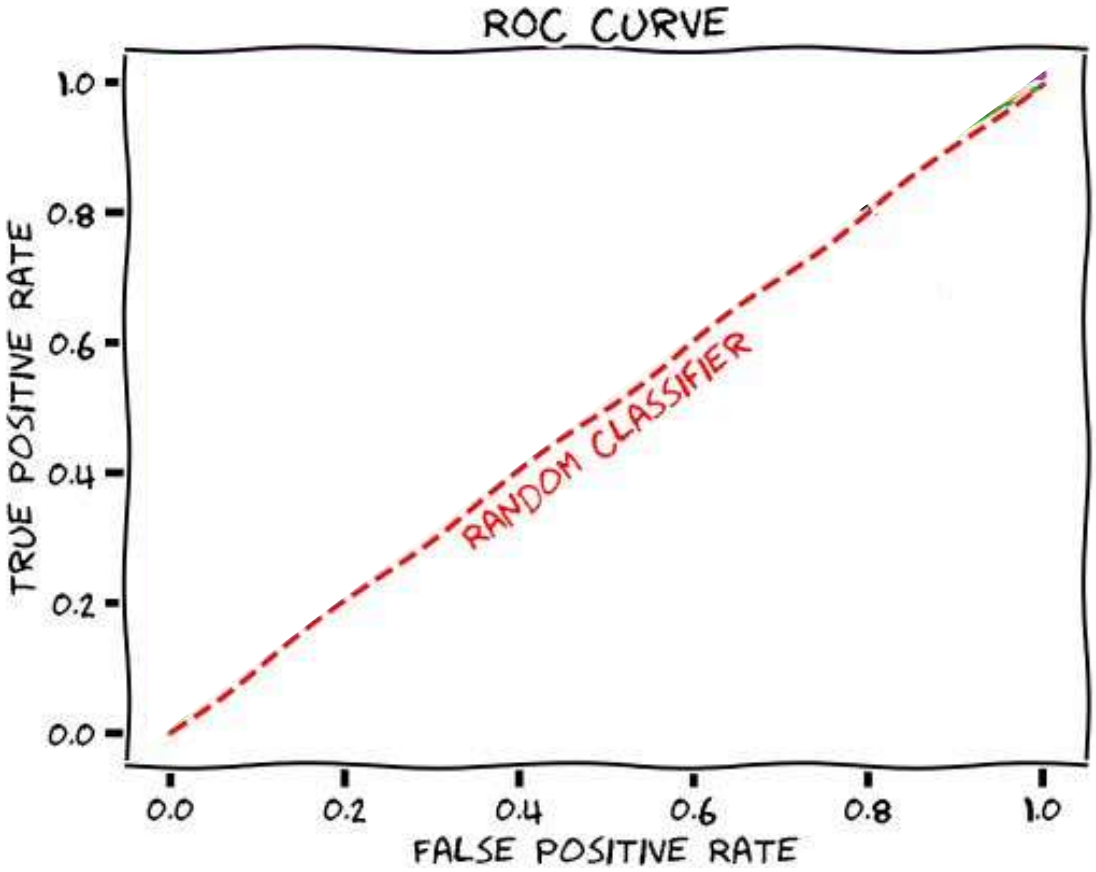
		Predicted condition			
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** hit	**False negative (FN),** type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	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 $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN}$ $= 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

Accuracy Metrics



(Glass Box: Measuring Performance, Retrieved 2021)

Accuracy Metrics

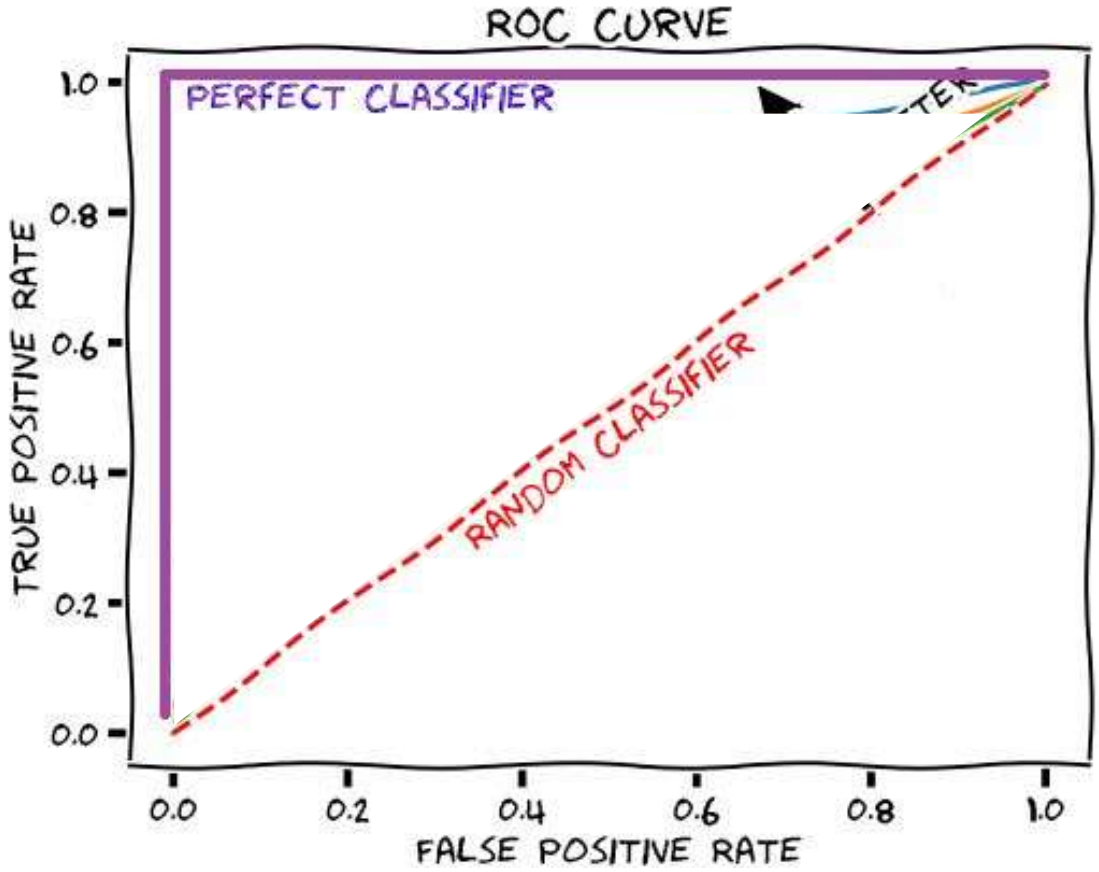


Logistic regression gives us a probability which we threshold (often at 0.5) to get predicted true(1) versus predicted false(0).

Change FP and TP rate by varying the threshold for predicting true or false.

(Glass Box: Measuring Performance, Retrieved 2021)

Accuracy Metrics

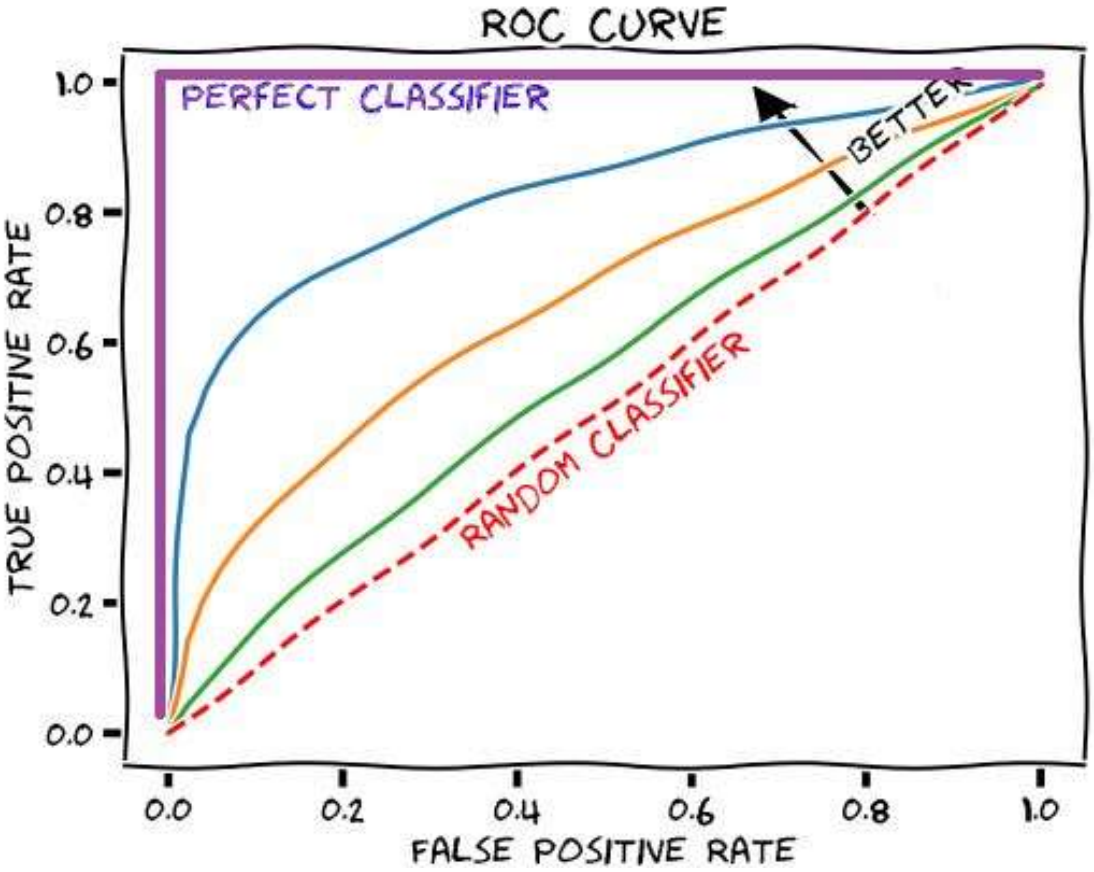


Logistic regression gives us a probability which we threshold (often at 0.5) to get predicted true(1) versus predicted false(0).

Change FP and TP rate by varying the threshold for predicting true or false.

(Glass Box: Measuring Performance, Retrieved 2021)

Accuracy Metrics



Logistic regression gives us a probability which we threshold (often at 0.5) to get predicted true(1) versus predicted false(0).

Change FP and TP rate by varying the threshold for predicting true or false.

(Glass Box: Measuring Performance, Retrieved 2021)

Mediation Analysis

Path Analyses (a type of “structured equation modeling”)

How much does **M** mediate the effect of
X on **Y**?

$$Y = \beta_{0c'b} + c'X + bM + \epsilon_{c'b}$$

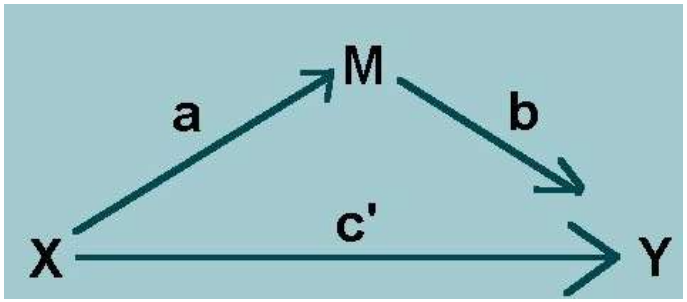
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

Mediation Analysis

Path Analyses (a type of “structured equation modeling”)

How much does **M** mediate the effect of **X** on **Y**?

$$Y = \beta_{0c} + cX + \epsilon_c$$
$$X = \beta_{0a} + aM + \epsilon_a$$
$$Y = \beta_{0c'b} + c'X + bM + \epsilon_{c'b}$$



(Kenney, 2015)

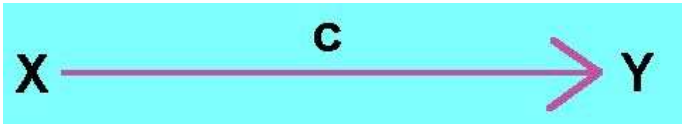
http://davidakenny.net/cm/mediate.htm

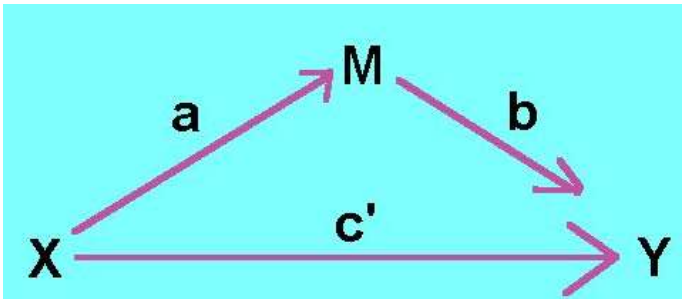
Mediation Analysis

Path Analyses (a type of “structured equation modeling”)

How much does **M** mediate the effect of **X** on **Y**?

$$Y = \beta_{0c} + cX + \epsilon_c$$
$$X = \beta_{0a} + aM + \epsilon_a$$
$$Y = \beta_{0c'b} + c'X + bM + \epsilon_{c'b}$$





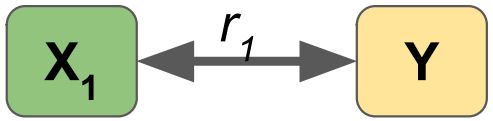
Effect size: often reported as c - c'.

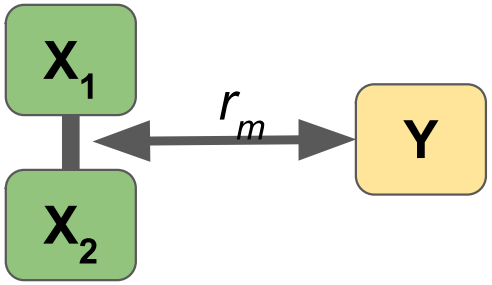
Used for *basic* causal inference.

(Kenney, 2015)

http://davidakenny.net/cm/mediate.htm

Moderation (interaction)



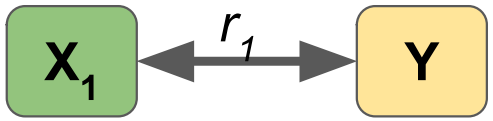


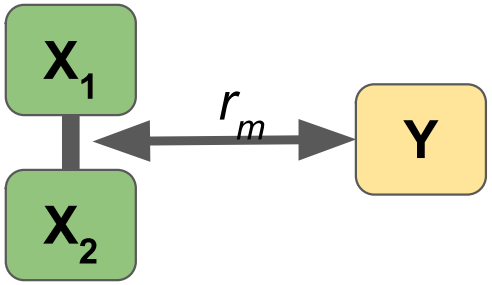
When $r_1 \neq r_m$, X_2 moderates the relationship between X_1 and Y.

Examples:

?

Moderation (interaction)





When $r_1 \neq r_m$, X_2 moderates the relationship between X_1 and Y.

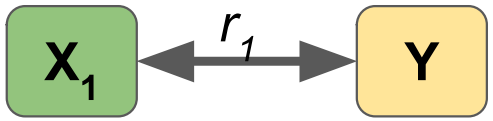
Examples:

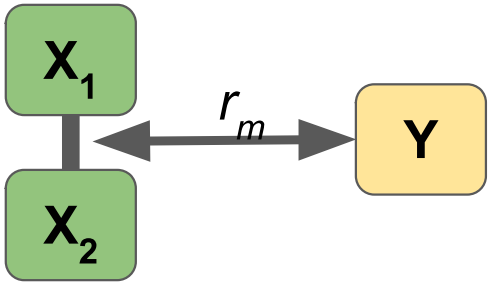
Y: Attend church? X_1: Agreeableness, X_2: From US?

Movie Reviews:

Y: Rated Depressing, X_1: “death” in review, X_2: Silly Horror Movie?

Moderation (interaction)





When $r_1 \neq r_m$, X_2 moderates the relationship between X_1 and Y.

More precisely moderation analyses fit the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_M X_1 X_2 + \beta_2 X_2 + \epsilon$$

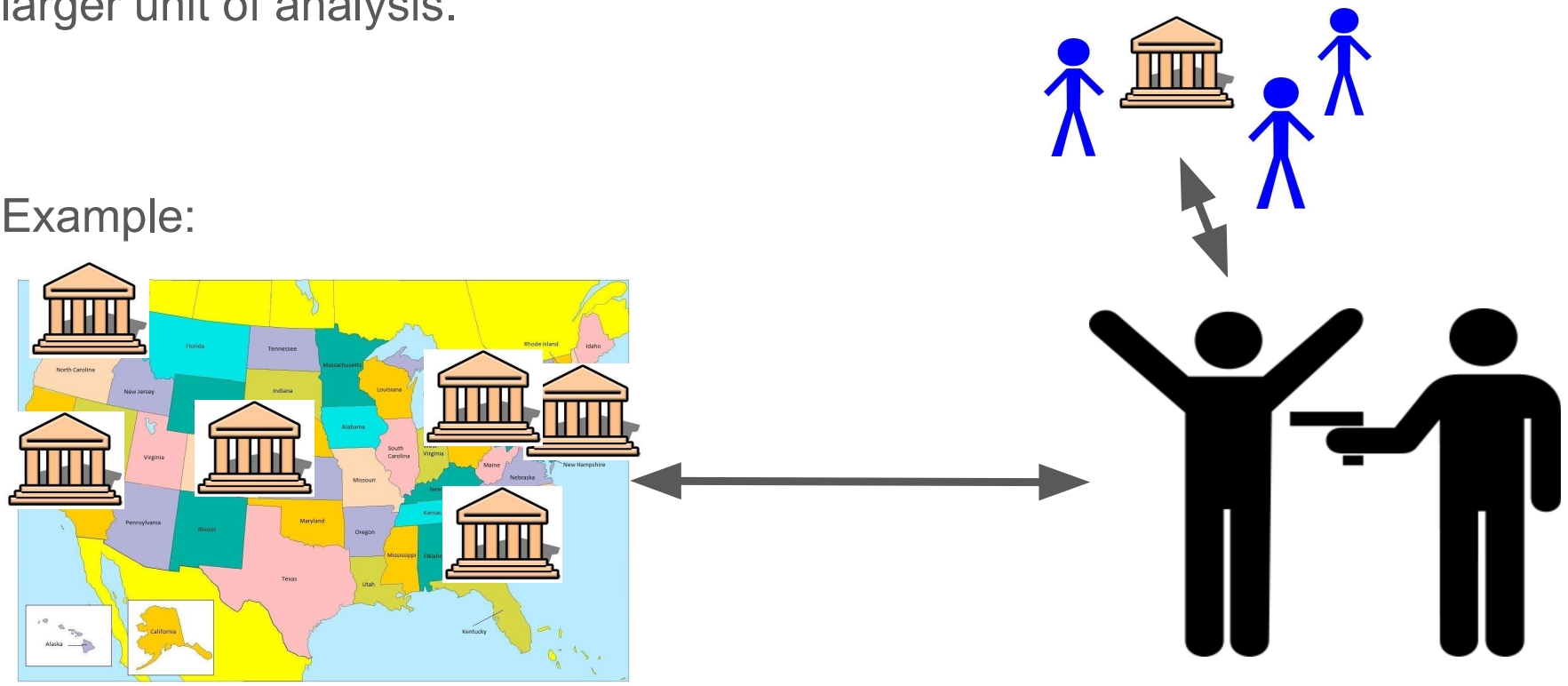
$X_1 X_2$: The interaction term. (Element-wise multiplication)

β_M can then be tested for significance using the same t-test we use for any individual coefficient in multiple linear regression

Ecological Fallacy

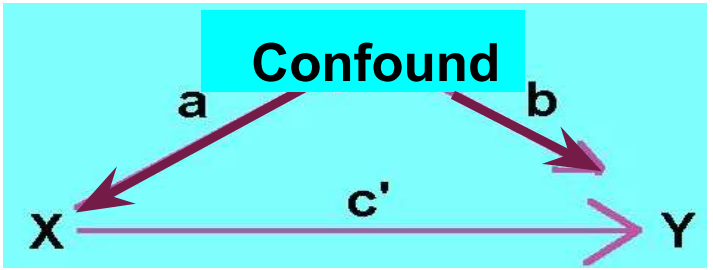
The assumption that an effect at one unit of analysis will hold for a smaller or larger unit of analysis.

Example:



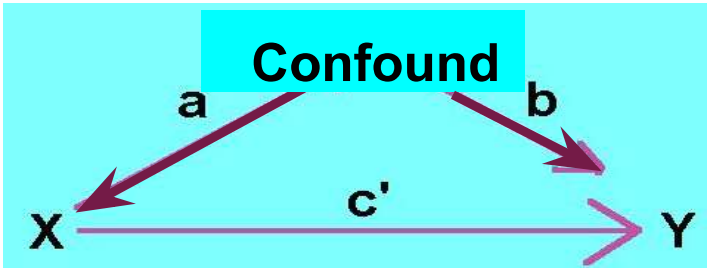
Multi-Level Models

Problem: Sometimes variables at one unit of analysis are *confounded* by a variable at another level.



Multi-Level Models

Problem: Sometimes variables at one unit of analysis are *confounded* by a variable at another level.



Examples:

Pot heads are more likely to say “hella”
but really californians are more like to say “hella” and be potheads.

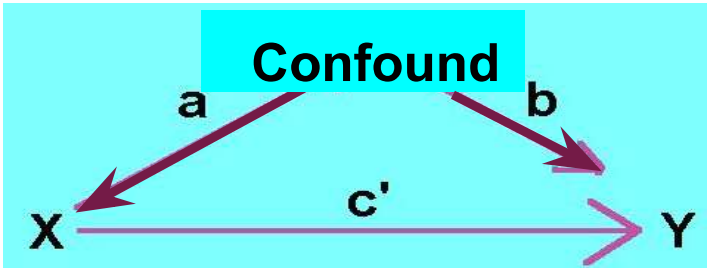
X = use of “hella”

Y = pot-head or not

Confound = from california?

Multi-Level Models

Problem: Sometimes variables at one unit of analysis are *confounded* by a variable at another level.



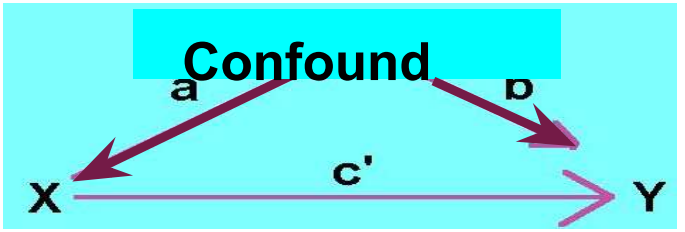
Examples:

Pot heads are more likely to say “hella”
but really californians are more like to say “hella” and be potheads.

Females are more likely to post pictures of food
but really both food posts and females are more common on Pinterest.

Multi-Level Models

Problem: Sometimes variables at one unit of analysis are *confounded* by a variable at another level.



Examples:

Pot heads are more likely to say “hella”
but really californians are more like to say “hella” and be potheads.

Females are more likely to post pictures of food
but really both food posts and females are more common on Pinterest.

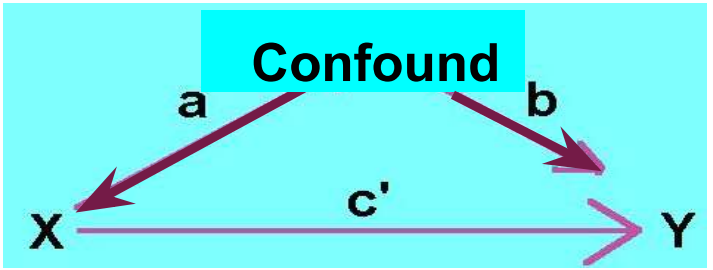
Solution: include aggregate confounding variable as a covariate in multiple linear regression. (also useful for prediction)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

Multi-Level Models

Problem: Sometimes variables at one unit of analysis are *confounded* by a variable at another level.

$$Y = \beta_0 + \beta_1 X_1 + \beta_A A + \epsilon$$



A : aggregate indicator variable (is in region or not? Pinterest usage).

Solution: include aggregate confounding variable as a covariate in multiple linear regression. (also useful for prediction)

Fixed-Effects Model (a multi-level model useful for time data)

$$y_{it} - \overline{y}_i = \left(X_{it} - \overline{X}_i\right)\beta + (\alpha_i - \overline{\alpha}_i) + (u_{it} - \overline{u}_i) \implies \ddot{y}_{it} = \ddot{X}_{it}\beta + \ddot{u}_{it}$$

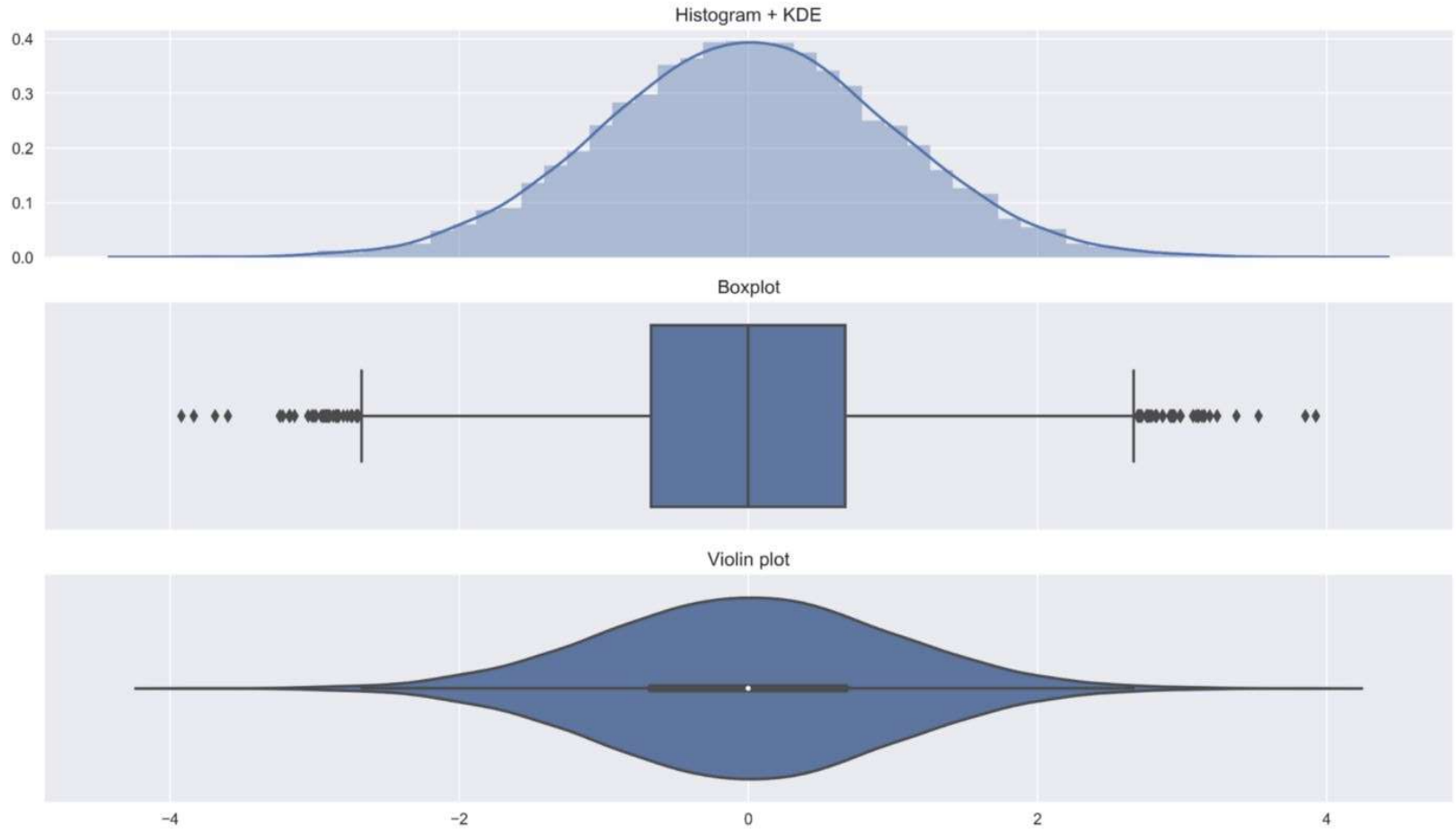
$$\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it} \qquad \overline{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it} \qquad \overline{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$$

$$y_{it} = X_{it}\beta + \alpha_i + u_{it}$$

- y_{it} is the dependent variable observed for individual i at time t.
- X_{it} is the time-variant $1 \times k$ (the number of independent variables) regressor vector.
- β is the $k \times 1$ matrix of parameters.
- α_i is the unobserved time-invariant individual effect. For example, the innate ability for individuals or historical and institutional factors for countries.
- u_{it} is the error term.

Unlike X_{it}, α_i cannot be directly observed.

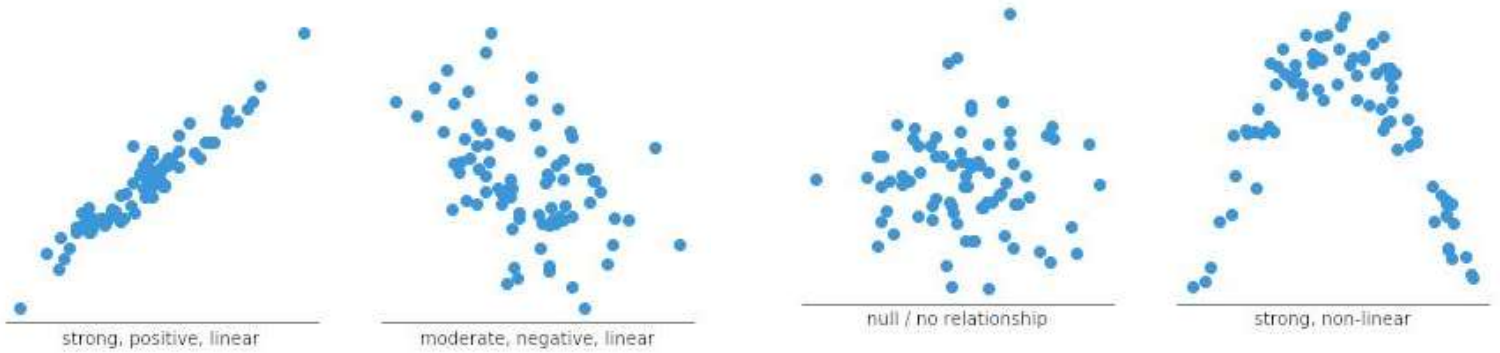
Useful Plots: For distributions



(Lewinson, 2019)

Useful Plots: Correlation

Scatter Plot: for two variables expected to be associated (with optional regression line)



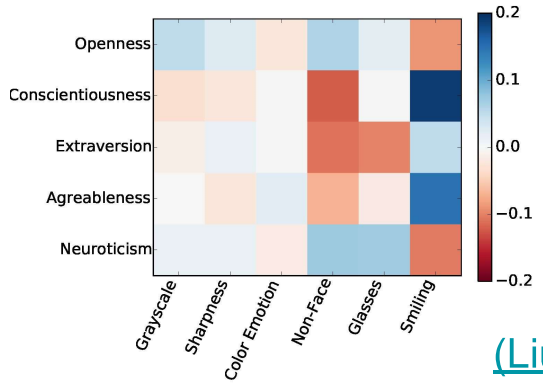
(Chartio)

Correlation Matrix: for comparing associations between many variables (use Bonferroni correction if hyp testing)

	FriendSize	Intelligence Quotient	Income	Sat W/ Life	Depression
F1	0.03	0.04	0.12	0.02	-0.1
F2	0.04	-0.26	-0.19	-0.09	0.11
F3	-0.07	-0.13	0.02	-0.02	-0.02
F4	-0.03	0.27	-0.08	-0.12	0.11
F5	-0.01	0.23	0.29	0.07	-0.21

Fig 3. Individual factor correlations with outcomes. Note how F4 which captures the use of swear words negatively correlates with Satisfaction with Life (SWL).

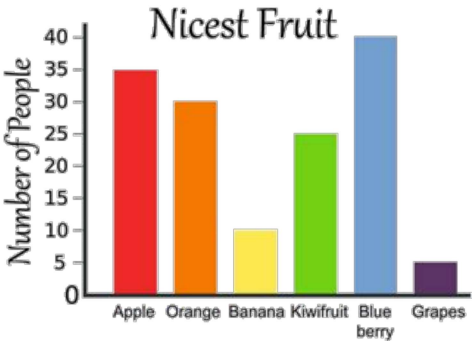
https://doi.org/10.1371/journal.pone.0201703.g003



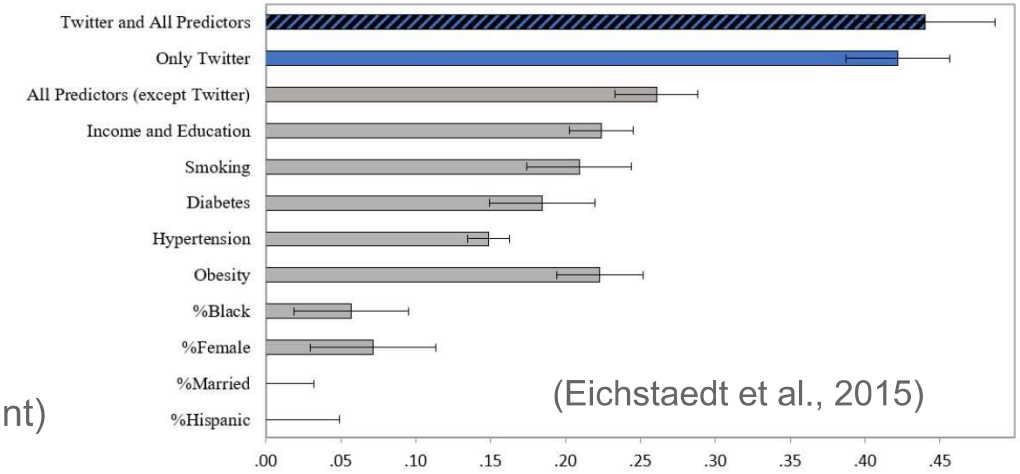
(Liu et al., 2016)

Useful Plots: Any Values

Bar Plot: To visually compare values under multiple conditions.

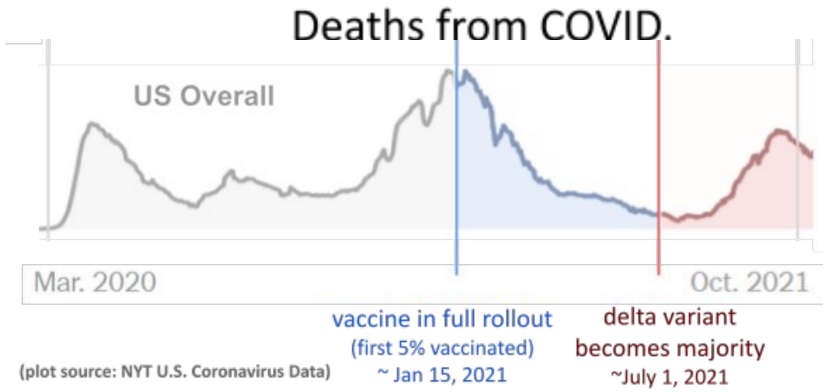


(Science sEDiment)



(Eichstaedt et al., 2015)

Line Plot: When one variable has a natural ordering (e.g. time)



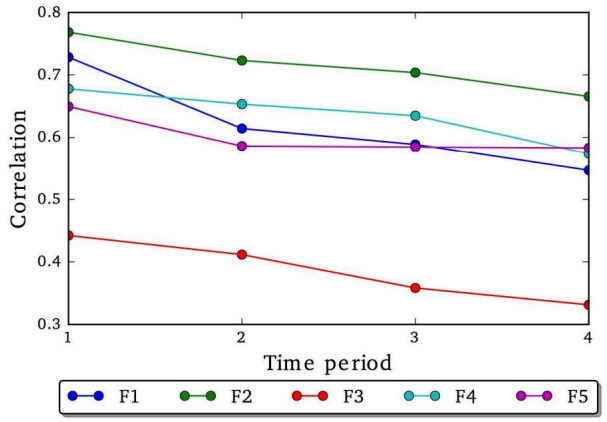


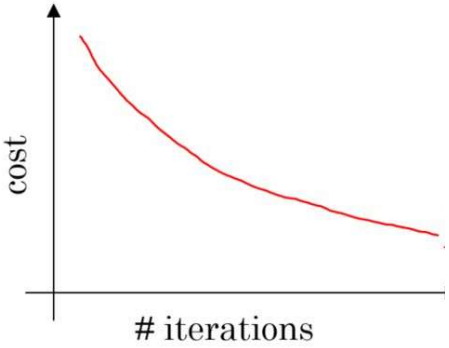
Fig 6. Test re-test validity of our learned factors.

https://doi.org/10.1371/journal.pone.0201703.g006

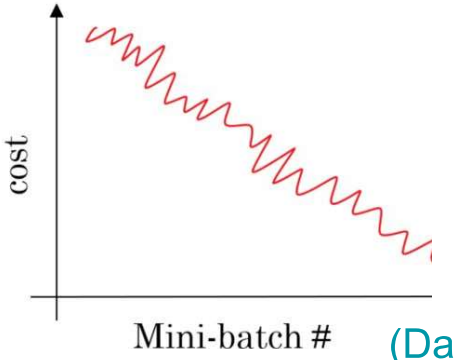
Useful Plots: Prediction

Learning Curve: for plotting error from gradient descent.

for a model with convex optimization (i.e. linear regression)

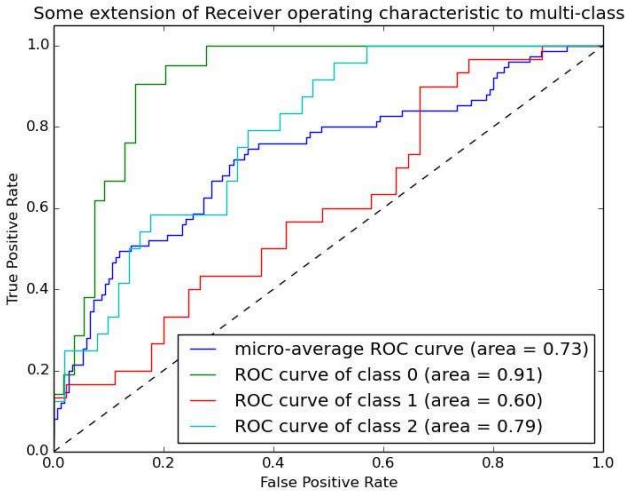


for a model with non-convex optimization (i.e. most deep learning)

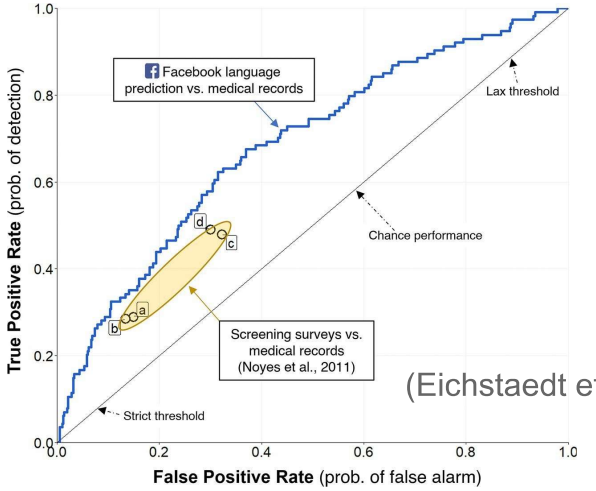


(Dabura, 2017)

ROC Plot: for visualizing true-positive to false-positive rates (used for AUC metric)



(PLOT_ROC)



(Eichstaedt et al., 2018)