Regression

ehsan.karim@ubc.ca

Sept 24, 2020

SPPH 504/007

What was the most difficult part of week 2

- Following video materials
- Using wall of confusion
- Coming to office hour
- Quiz questions
- Concept questions
- Lab exercises
- Thinking about final project
- 2nd wave of pandemic is coming?

Reference

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2011) [chapter 10] "**Predictor Selection**". In: **Regression methods in biostatistics**: linear, logistic, survival, and repeated measures models. Springer.

(available in the "Library Online Course Reserves": see the Canvas link on the left).

Reference

Greenland, S., & Pearce, N. (2015). Statistical foundations for model-based adjustments. Annual review of public health, 36, 89-108.

Inferential goals

Page 396

1. Prediction
2. Evaluating a predictor of primary interest
3. Identifying the important independent predictors of an outcome
4. Descriptive (?)

Inferential goals

Example 1

Association Between Use of Interferon Beta and Progression of Disability in Patients With Relapsing-Remitting Multiple Sclerosis

Afsaneh Shirani, MD; Yinshan Zhao, PhD; Mohammad Ehsanul Karim, MSc; et al

» Author Affiliations | Article Information

JAMA. 2012;308(3):247-256. doi:10.1001/jama.2012.7625

Abstract

Context Interferon beta is widely prescribed to treat multiple sclerosis (MS); however, its relationship with disability progression has yet to be established.

Objective To investigate the association between interferon beta exposure and disability progression in patients with relapsing-remitting MS.

Inferential goals

Example 2

Development and Validation of a Prognostic Index for 1-Year Mortality in Older Adults After Hospitalization

Louise C. Walter, MD; Richard J. Brand, PhD; Steven R. Counsell, MD; et al

Abstract

Context For many elderly patients, an acute medical illness requiring hospitalization is followed by a progressive decline, resulting in high rates of mortality in this population during the year following discharge. However, few prognostic indices have focused on predicting posthospital mortality in older adults.

Objective To develop and validate a prognostic index for 1 year mortality of older adults after hospital discharge using information readily available at discharge.

Goal 1: Goal of prediction models

Page 396/7/8

"**Prediction error** (PE) measures how well the model is able to predict the outcome for <u>new observations</u> not used in developing the prediction model."

- Bias reduced for models with more variables
- Unimportant variables lead to noise / variability
- Bias variance trade-off / need penalization

What is the difference between R-squared (R2) and the adjusted R2

R2 and adjusted R2 increases with every predictor added to a model

Lowest value of R2 and adjusted R2 is zero

Adjusted R-square penalizes you for adding variables which do not improve your existing model.

More unimportant variables you add into the model, the gap in R2 and adjusted R2 increases.

Goal 1: Measures of PE

Page 397

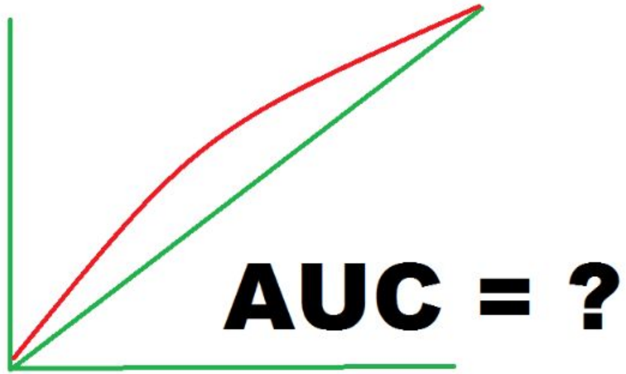
Continuous

- R-squared
- Adjusted R-squared

Binary

- Brier score, Brier score scaled
- Nagelkerke's R-squared (glm)

What is the Area Under Curve (AUC) value here?



AUC = 1

AUC = 0

AUC = 0.5

AUC = 0.6

AUC = 0.95

AUC = 10

Discrimination and calibration

Discrimination (how well prediction model can discriminate Y=0 vs Y=1)

- AUC from ROC / C-statistics
  - C-stat = 0.98 ~ Nagelkerke's R-square = 87%
  - C-stat = 0.7 - 0.8 ~ Nagelkerke's R-square = 10 - 20%

Calibration (agreement between obs vs. predicted)

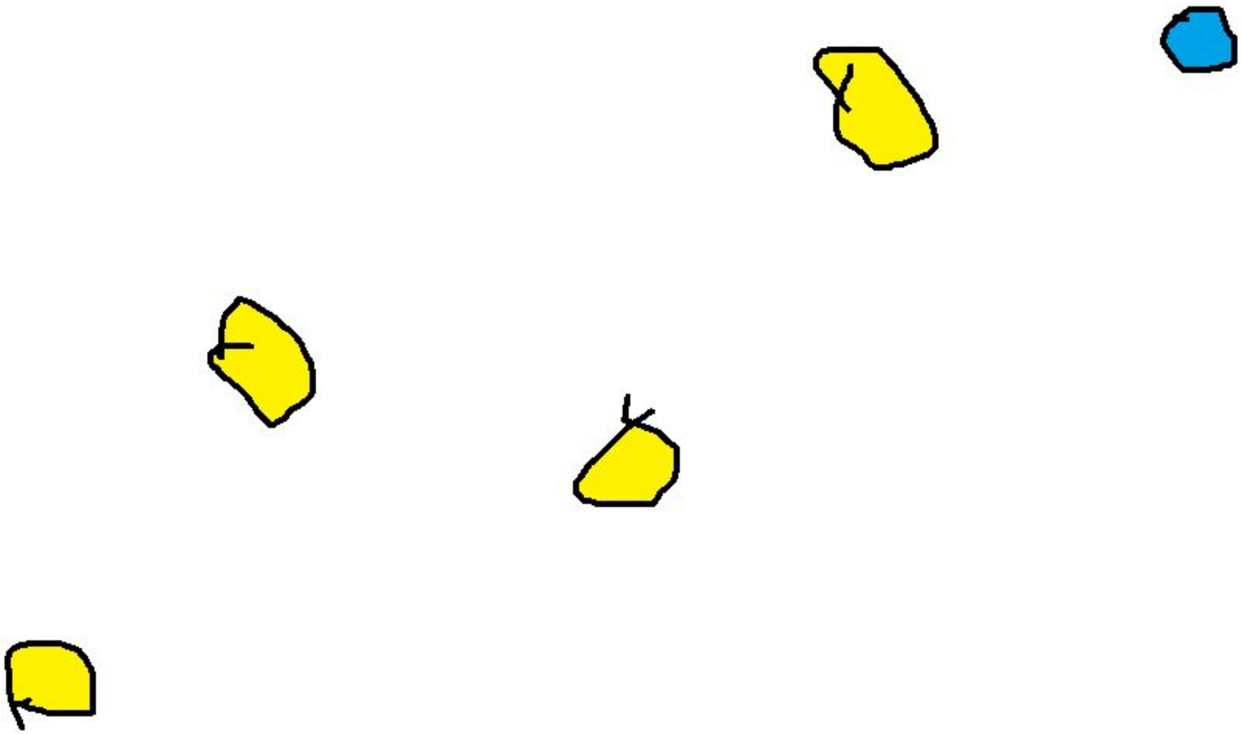
- Hosmer-Lemeshow test

Goal 1: Overfitting / Optimism

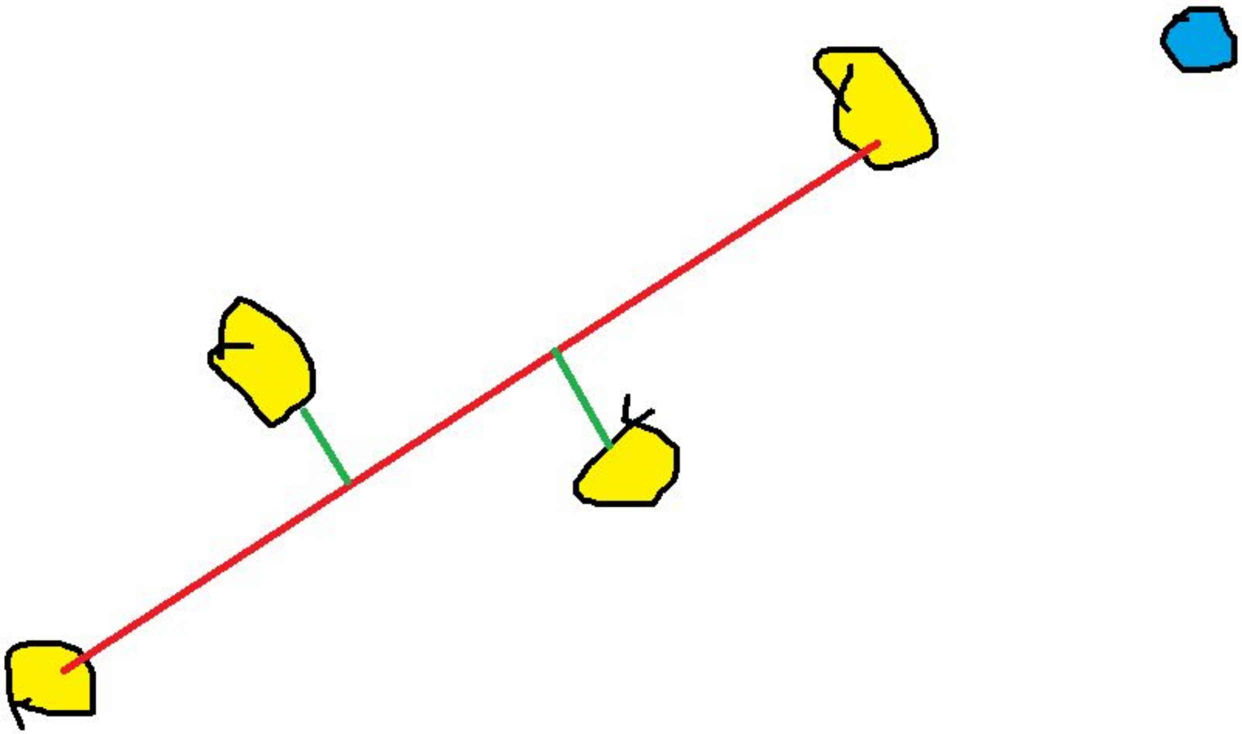
Page 397

- Population > Sample (empirical data)
- Predictive model built on empirical data
- Model performs very well in the empirical data where the model was fitted (optimistic)
- Model performs poorly in the new data (generalization is not as good)

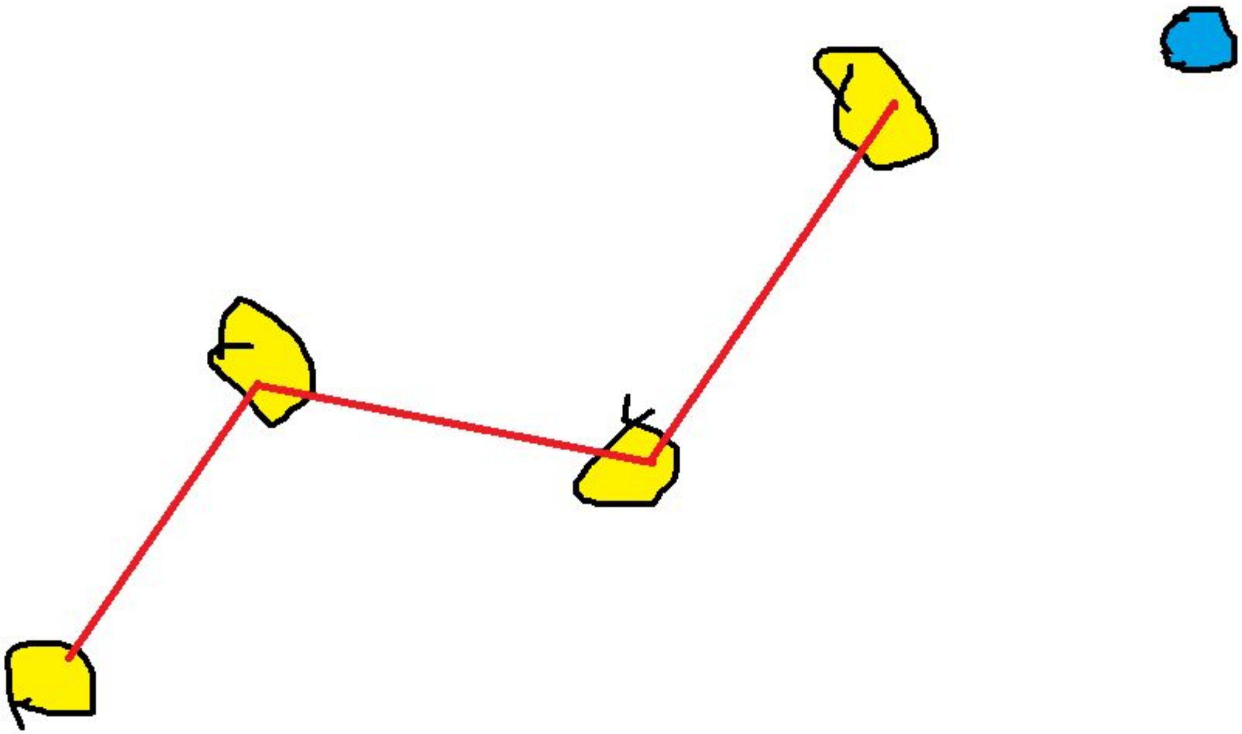
Goal 1: Overfitting / Optimism



Goal 1: Overfitting / Optimism



Goal 1: Overfitting / Optimism



Goal 1: Overfitting / Optimism

Causes

- Model determined by data at hand without expert opinion
- Too many model parameters (age, age^2, age^3) / predictors
- Too small dataset (training) / data too noisy

Consequences

- Overestimation of effects of predictors
- Reduction in model performance in new observations

How to validate model (reduce optimism)

1. Internal validation

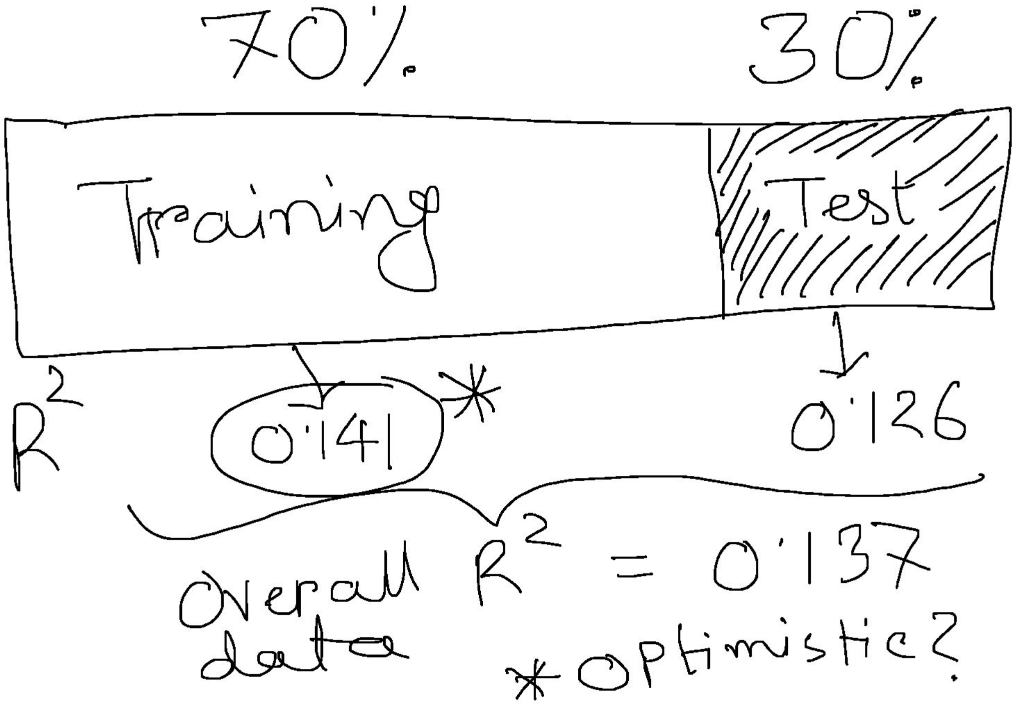
- Apparent validation (100% data; stable; optimistic; used as a reference)
- Split sample
- Cross-validation (CV), Leave-one-out CV
- Bootstrap, .632 and .632+ bootstrap

2. External validation

- Temporal
- Geographical
- Different data source to calculate same variable
- Different disease

Goal 1: Optimism-Corrected PE

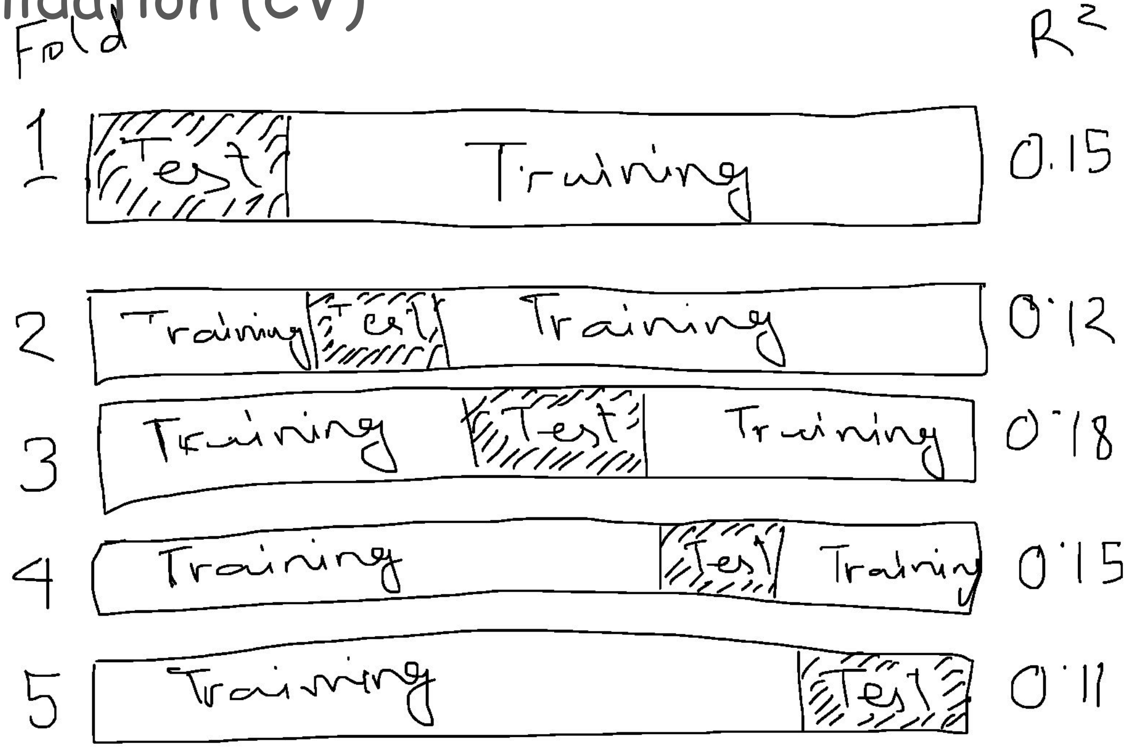
1. Split sample approach



Goal 1: Optimism-Corrected PE

Page 399

2. K fold Cross-validation (CV)



Goal 1: Optimism-Corrected PE

3. Bootstrap

Actual data

ID	ch	Tri
1	362	209
2	194	79
3	81	165

$\overline{ch} = 212.3$

Bootstrap sample 1

ID	ch	Tri
1	362	209
3	81	165
3	81	165

$\overline{ch} = 174.7$

Bootstrap sample 2

ID	ch	Tri
2	194	79
2	194	79
1	362	209

$\overline{ch} = 250$

Goal 1: selection the model

1. Pre-specify variables
    a. based on subject-area knowledge / expert-opinion / meta-analysis
2. Use m/10 or m/20 rule where m = effective sample size (# of obs.) to identify candidate predictors.
    a. 10 or 20 obs per predictor without looking at the outcome (no data peeking)
3. Use CV to do model selection
    a. based on r-squared/AIC/BIC
    b. Event per variable is another concern
4. Use shrinkage method
    a. These are useful for collinearity reduction (will learn later)
    b. Alternatively use CV / bootstrap to decide if a collinear variable is to keep / delete
    c. Generally other than extreme scenarios, would try to include if PE is reduced after inclusion

Collinearity

How to identify?

Collinearity

$$\hat{y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

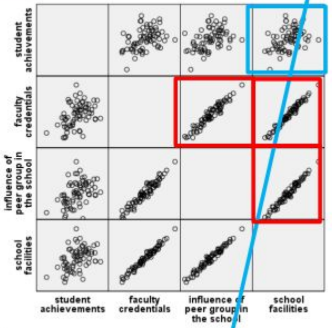
Estimated coefficient of a variable has an **opposite sign** from what you would expect from correlation table.

The predictor variables are so **highly correlated** that each one may serve as a proxy for the others in the regression equation **without affecting the total explanatory power**.

Correlations

		student achievements	faculty credentials	influence of peer group in the school	school facilities
student achievements	Pearson Correlation	1	.419**	.440**	.418**
	Sig. (2-tailed)		.000	.000	.000
	N	70	70	70	70
faculty credentials	Pearson Correlation	.419**	1	.960**	.986**
	Sig. (2-tailed)	.000		.000	.000
	N	70	70	70	70
influence of peer group in the school	Pearson Correlation	.440**	.960**	1	.982**
	Sig. (2-tailed)	.000	.000		.000
	N	70	70	70	70
school facilities	Pearson Correlation	.418**	.986**	.982**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	70	70	70	70

**. Correlation is significant at the 0.01 level (2-tailed).



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	73.506	3	24.502	5.717	.002[b]
	Residual	282.873	66	4.286		
	Total	356.379	69			

Contradictory F-ratio and t-statistics.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	-.070	.251		-.279	.781	-.570	.430
	faculty credentials	1.101	1.411	.525	.781	.438	-1.715	3.918
	influence of peer group in the school	2.322	1.481	.945	1.568	.122	-.635	5.280
	school facilities	-2.281	2.220	-1.027	-1.027	.308	-6.714	2.152

a. Dependent Variable: student achievements

These are typical of a situation where **extreme collinearity** is present.

Page 421

When VIF is greater than 20, we should

Delete at least one collinear variable with high VIF

Do nothing

Combine all collinear variables to a single variable

Goal 1: Why do we need predictor selection

- Too large model may be hard/impractical to deal with computationally
- Some predictors may be really irrelevant/unimportant/implausible to have any effect

Goal 1: Wrong reasons to omit X (prediction)

- Insignificant
  - prediction is about estimation, not hypothesis testing
- Collinearity
  - Fearing instability
- Model parsimony
  - Simpler explanation = simplistic model

Goal 2: Primarily interested in Y-A relationship

- Adjust for everything?
  - Empirical Criteria:
    - pre-treatment,
    - common cause,
    - disjunctive,
    - modified disjunctive,
    - modified modified disjunctive
  - Data sparsity:
    - variance inflation
  - Multicollinearity
    - variance inflation
    - Could be fine if SE is not too high
      - Use bootstrap to see how unstable the results are.
- See S Greenland, N Pearce (2015) [p96]

Page 407

Goal 2: Primarily interested in Y-A relationship

Exclude the following variables

- Alternative measures of outcome
- Alternative measures of exposure
- Some variables based on DAG knowledge
  - Mediator
  - Known instrument from the literature
  - Effect of outcome

Goal 2: How to select covariates?

1. Subject area knowledge
    a. DAG
    b. Vanderwalee paper from previous pre-reading for some practical guidance
2. Statistical ground
    a. Best subset
    b. Stepwise / forward
    c. Backward elimination (BE)
    d. Bivariate screening (a variant of BE) - either omit or use larger cut-point (e.g., 0.5)
    e. Bootstrap on selection (all predictors selected via BE in 50% of the bootstrap samples)
3. Interaction / effect modification are part of model specification

Parsimony versus Confounding

A worthwhile task for goal 2?

- Probably not
- Precision gain is often argued, but that gain from variable selection might be misleading
- Primary goal should be reduction of confounding
- Still a debatable issue
- See S Greenland, N Pearce (2015) [p99-100]

AIC based stepwise model selection comes up with different variables than p-value based stepwise model selection

TRUE

FALSE

Model selection

Page 425

- Smaller P-values / narrower CIs than the truth
- Post-selection bias / selective inference is a problem for goal 2 (causality)
    - Borderline p-values need to be assessed carefully
- Not much of a problem for goal 1 (prediction)?
    - as long as CV is properly used (as per text).

Stepwise / FS / BE

Advantages

- Easy to implement / objective

Disadvantages

- Instability in selection
- Biased estimation of ultimately selected coefficients
- Selective inference
- P-value of 0.157 ~ AIC
  - Almost similar criterion different cutpoint

Goal 3: Identify important predictors

- Still need to deal with confounding
  - More complicated DAG
- Variable importance (will learn later)

General Issues: Centering and scaling

- diastolic blood pressure = 0 (no pressure)
  - lab 1a deals with this issue
  - May be centered to what is clinically considered as normal (say, 80)
- Age in 1 year has clinical impact on chronic disease?
  - Consider scaling to 10 years

See S Greenland, N Pearce (2015) [p 93-94]

General Issues: Collapsibility

Change-in-Estimate Strategies in OR. Is this a problem?

S Greenland, N Pearce (2015): [p98]

- “CIE methods have an advantage over selection based only on outcome or exposure prediction insofar as the selection criterion is on the scale used for contextual interpretation.”
- See lab 3 C
- may not be a problem for rare disease

Thanks!

ehsan.karim@ubc.ca