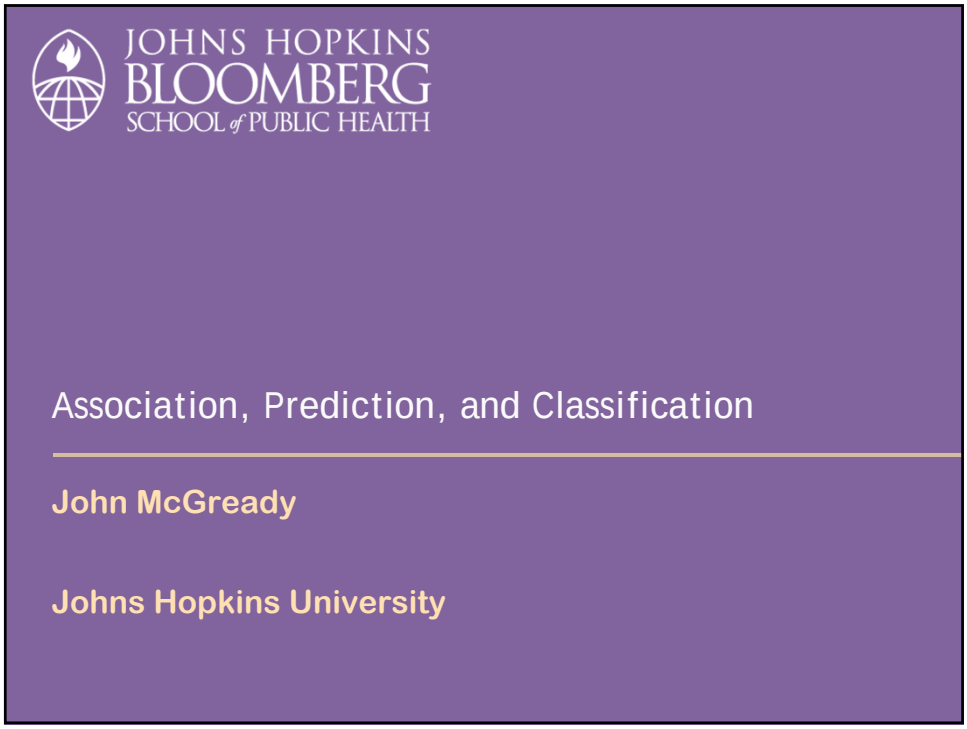
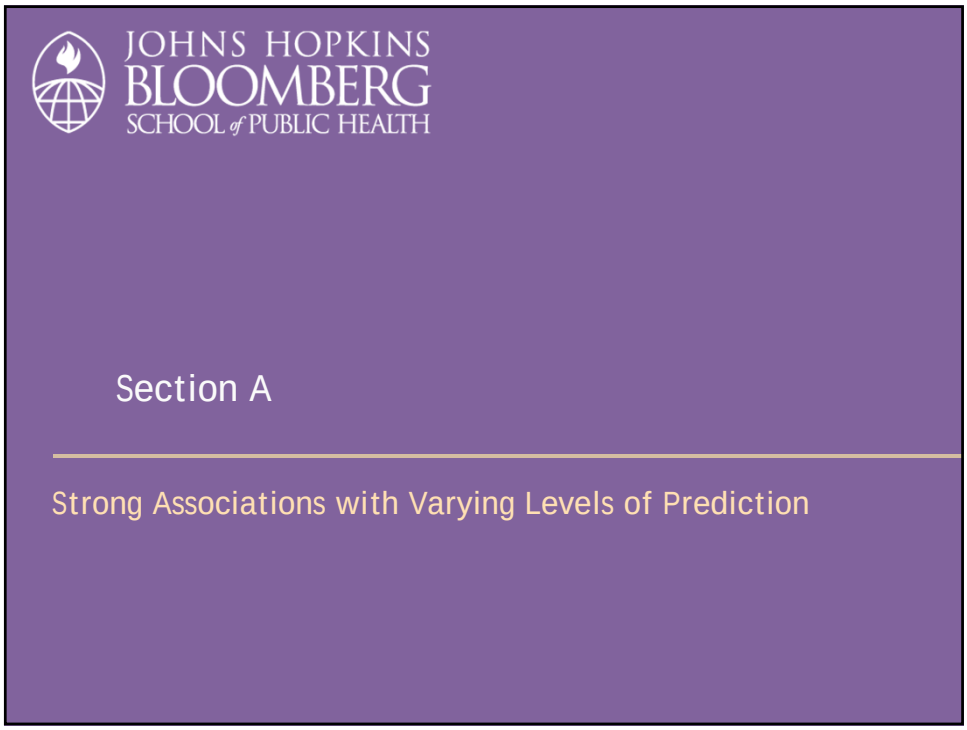

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH
Association, Prediction, and Classification
John McGready
Johns Hopkins University


JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH
Section A
Strong Associations with Varying Levels of Prediction

Associations

- Statistical associations and trends are important in medicine and public health, but do not necessarily translate into good individual level associations
- Statistically significant differences show fundamental shifts in distributions of individual values between groups (mean shifts up by 5 mmHg for women on OCs, proportion of infants born with HIV shifts down by 15% when mother is treated with AZT during pregnancy). However, there may still be substantial crossover in individual values between the groups.

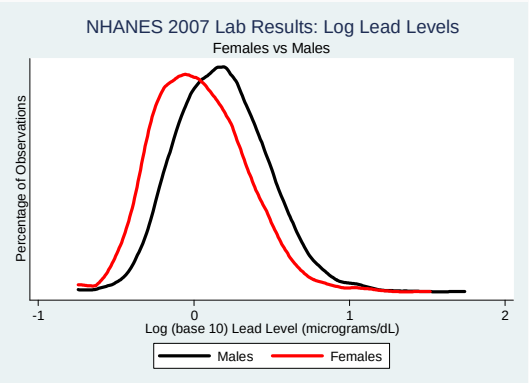
3

Associations: Example 1

- Lead Levels, Females and Males from US: Strong Association, Low Predictive Ability

M: 0.04

F: 0.19



Difference in Means: 0.145 (95% CI: 0.13- 0.16), $p < 0.0001$

4

Associations: Example 2

- HIV Mother to Infant Transmission Study

18 month transmission percentages:

AZT: 0.07 (95% CI: .04 to 0.12) Placebo: 0.22 (95% CI 0.16 to 0.28)

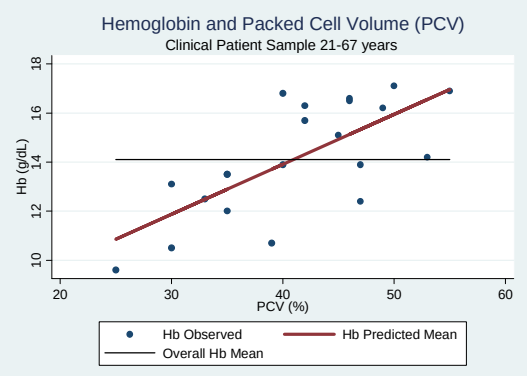
Difference: (AZT to Placebo): -0.15 (-0.08 to -0.22)

AZT is substantially, and statistically significantly associated with a reduction in maternal/infant HIV transmission; but can we tell which mother/infant pairs will benefit from AZT in future treatments?

5

Associations: Example 3

- Hemoglobin and Packed Cell Volume: clinic sample of 21 patients

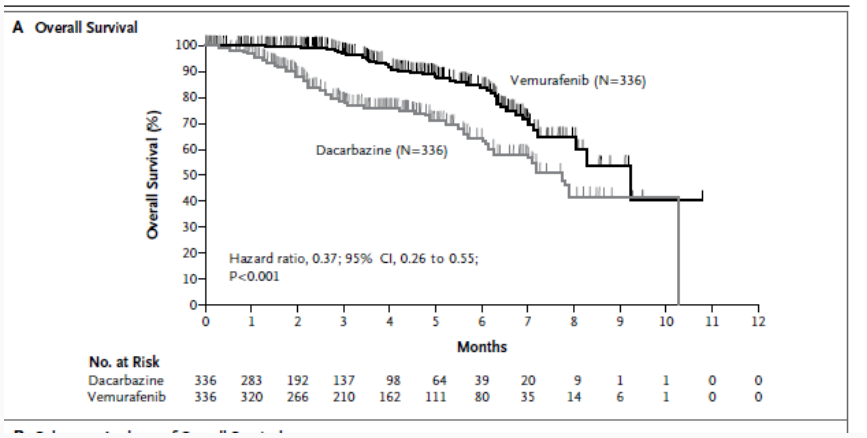


Regression slope: 0.20 (95% CI: 0.10 to 0.30), p< 0.0001

6

Association: Example 4

- RCT for Melanoma treatments



7

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section B

Model Based Prediction: Assessment

Prediction

- Predictive models tend to built using multivariable models: as some of the examples demonstrated in section A, it's hard to get good prediction of an outcome using a single predictor
- Multivariable regressions: linear, logistic, Cox are all tools for building models that give magnitude and significance of the association between each predictor and the outcome, and can be evaluated in terms of the model's ability to predict the outcome
- Assessments of prediction involve some comparison of the discrepancy between of the observed outcomes and the predicted outcomes

9

Uses of prediction models

- Clinical decision making
- ID high risk persons for preventive interventions
- ID high risk persons for clinical or epi studies
- Medical/biologic insight
- Risk information might be useful to a patient/family for planning purposes
- Predicting presence of risk factors (!), e.g. BRCA1.

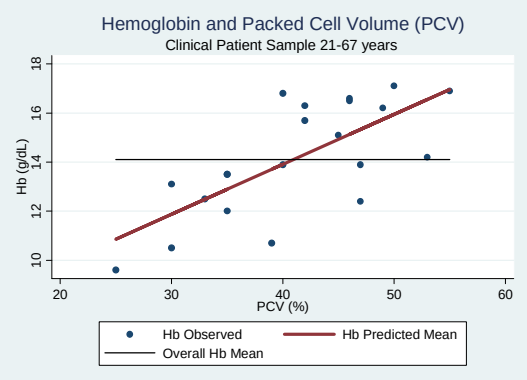
Linear Regression

- Often used measure of predictive power of a linear regression model is call R^2
- R^2 how much of the original, individual level variability in an outcome is explained by taking predictor(s) information into account

11

Prediction Assessment: Linear Regression

- Hemoglobin and Packed Cell Volume: estimated R^2=0.51



Roughly 51% of the subject's variation in Hg is explained by taking their Hg into account.

12

Prediction Assessment: Linear Regression

- Hemoglobin and Packed Cell Volume, and Age

If a multivariable regression model relating HB to PCV and age is employed, R^2 increases to 82%.

- R^2 gives information about how well model predicts at individual level, above and beyond associations at the populations level
- R2 suffers from some "issues"
  - R2 will automatically increase with each additional predictor added whether this predictor is association

13

Prediction Assessment: Linear Regression

- R^2 suffers from some "issues"
  - R^2 will automatically increase with each additional predictor added whether this predictor adds information about outcome above and beyond other predictors: this is kept in check by another measure called "Adjusted R2"
  - R^2 is overly optimistic: a model generally fits the data it was estimated with better than the population from which the data was sampled: this is a problem with most model based measures of prediction, and we will discuss in more detail

14

Prediction Assessment: Logistic Regression

- Two Issues to be addressed when assessing prediction of binary outcomes with logistic regression
  - How well do the predicted probabilities from the logistic regression models match the observed probabilities
  - How well does the model predict the binary outcome above and beyond flipping a coin?
- Goodness of fit tests:
  - Compare observed proportion of outcomes measured across a group to the average predicted probabilities for members of a group
  - Most famous for logistic regression: Hosmer-Lemeshow
    Ho: model fits well
  - Model may not "fit" data very well, but can still have reasonable predictive power: data may fit the data well but have poor predictive power

15

Prediction Assessment: Logistic Regression

- (Relative) Prediction of "individual" outcomes – how well does model predict "yes/no" outcome for individuals given his/her predictors (x's) above and beyond flipping a coin to predict a "yes/no"?
- Recall death/sepsis example from yesterday: for example, the estimated *probability* of death for a 50 year old patient with sepsis who has history of alcohol, but is not in shock, not malnourished, and does not have infarction at the time of surgery was 0.16%
- Suppose we use 0.16 as a cutoff for all future patients admitted to the ICU with sepsis for triage purposes: patients whose x's predict a probability of 0.16 or greater will be considered likely to die
  - How well would this cut off predict death versus flipping a coin?

16

Prediction Assessment: Logistic Regression

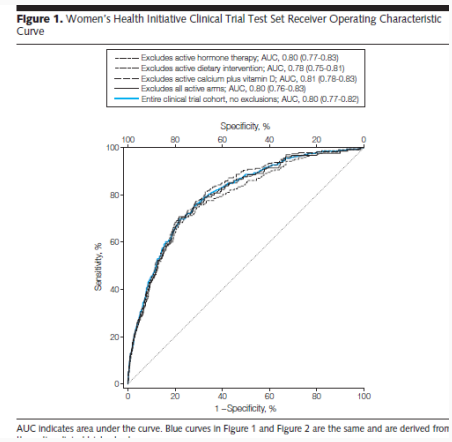
- Prediction of "individual" outcomes – how well does model predict "yes/no" outcome for individuals given his/her predictors (x's) above and beyond flipping a coin to predict a "yes/no"?

- Receiver-Operator Curve (ROC) analysis:
  - For each possible model predicted probability of being a yes/no:
  - Use the predicted probabilities for each individual in your sample to classify them as a "yes" of a "no" ;
  - compare these predicted classifications to the observed "yes's" or "no's:
  - Compute the observed sensitivity and specificity for the cutoff
  - Plot a curve of *sensitivity* (probability true positive) vs. *1-specificity* (this gives probability of false negative)
  - Area under curve gives "chance" of choosing outcome correctly

17

Prediction Assessment: Logistic Regression

- Ex: **Factors Associated With 5-Year Risk of Hip Fracture in Postmenopausal Women[1]:**



- [1] **Robbins J et al. Factors Associated With 5-Year Risk of Hip Fracture in Postmenopausal Women.** *Journal of American Medical Association.* **(2007) Vol 98, No 20.**

18

Prediction Assessment: Logistic Regression

- Issues with ROC curve assessment of predictive power of a logistic regression model: it is a relative prediction assessment tool, does not tell how well model classifies individuals

Internal Validity
- Overfitting: as with R^2 in linear regression models, area under ROC curve (AUC) tends to be better for data used to fit the model then for other samples from same population

External Validity
- Model will not likely predict as well for samples from populations different that the dataset used to fit the model

Clinical etc.. Utility
- So how does one use these results to classify subjects in future samples as "yes" or "no" for the outcome?

19

Example :Cox Regression

- Prediction can be assessed with Cox Regression as well

- Example: **The Effect of Including C-Reactive Protein in Cardiovascular Risk Prediction Models for Women[1]**

Background: While high-sensitivity C-reactive protein (hsCRP) is an independent predictor of cardiovascular risk, global risk prediction use.

Objective: To develop and compare global cardiovascular risk prediction models with and without hsCRP.

Conclusions: A global risk prediction model that includes hsCRPimproves cardiovascular risk classification in women, particularlyamong those with a 10-year risk of 5% to 20%. In models thatinclude age, blood pressure, and smoking status, hsCRP improvesprediction at least as much as do lipid measures.

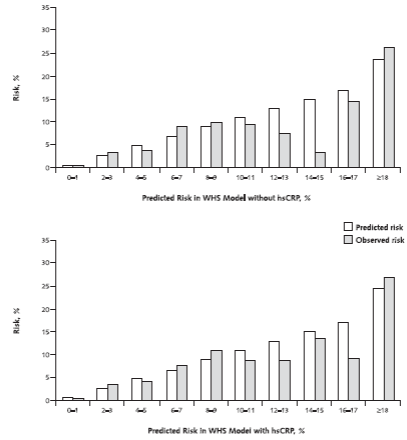
1 Cook N et al. The Effect of Including C-Reactive Protein in Cardiovascular Risk Prediction Models for Women. *Annals of Internal Medicine.* 2006;145:21-29.

20

Example :Cox Regression

- Presentation of observed outcomes versus predictions

Figure 3. **Calibration curves for risk prediction models without (*top*) and with (*bottom*) high-sensitivity C-reactive protein (*hsCRP*) in the model.**



The model that includes hsCRP shows closer agreement between observed and model-based predicted risk. WHS = Women's Health Study.

21

Example :Cox Regression

- Can also estimate a ROC curve for Cox regressions

22

Issues Common to All Prediction Assessments

- Internal Validity: Overfitting because model was estimated using a specific sample, and may not predict so well on other samples from sample population

Solution, yielding better estimate of predictive power in the population from which the sample was taken: Cross Validation

Randomly split sample into parts (usually 2/3 and 1/3 of entire sample)

Fit predictive model with first 2/3: called "Training sample"

Assess prediction on remaining 1/3: "test sample"

Data used to evaluation predictive power of the model was not used to estimated the model!

23

Issues Common to All Prediction Assessments

- Internal Validity: Overfitting because model was estimated using a specific sample, and may not predict so well on other samples from same population

Solution, yielding better estimate of predictive power in the population from which the sample was taken: Cross Validation

Randomly split sample into parts (usually 2/3 and 1/3 of entire sample)

Fit predictive model with first 2/3: called "Training sample"

Assess prediction on remaining 1/3: "test sample"

Data used to evaluation predictive power of the model was not used to estimated the model!

24

Issues Common to All Prediction Assessments

- External Validity: Does your sample represent the population of interest, and even if so, how well does model predict for samples from different population?

25

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section C

Model Based Prediction: Classification of Binary Outcomes

Classifying Subjects Based on Predicted Values

- Overall predictive power for logistic/Cox regressions correlates with increased ability to correctly classify subjects as to an outcome of interest: but overall assessment of predictive power not necessarily clinically useful
- As a clinician/scientist/outreach worker etc.. One may wish to:
  - Classify patients as "high risk" for disease and recommend further testing for those who are deemed "high risk"
  - Triage HIV- negative with high likelihood of contracting HIV to more intensive social services
  - Etc……
- Classification ability: the ability to predict for "future subjects" is a complex function of the relative predictive power of a model and the prevalence/incidence of an outcome

27

Classifying Subjects Based on Predicted Values

- How to best classify subjects given a predicted probability of an outcome?
- For example: you are using Pine's model to estimate risk of death in ICU for future admitted patients with sepsis using their age, alcohol history etc…

  Patient A has a predicted risk of death of 14%: do you triage him for more intensive services because he is "high risk" or not?

  How do you decide the cutoff for "high risk" versus "low risk", and how well does this cutoff discriminate?

28

Classifying Subjects Based on Predicted Values

- How to best classify subjects given a predicted probability of an outcome?
- The "best classification" cutoff/rule depends on your criteria: Do you wish to choose a cutoff to :
  - Maximize sensitivity OR
  - Maximize specificity OR
  - Maximize positive predictive value OR
  - Minimize overall classification error

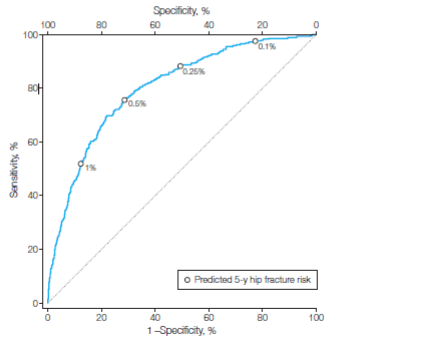
29

Classifying Subjects Based on Predicted Values

- From Hip Fracture Example: ROC curve with some cutoff choices labeled

RISK FACTORS ASSOCIATED WITH HIP FRACTURE IN POSTMENOPAUSAL WOMEN

Figure 2. Sensitivity and 1-Specificity of Receiver Operating Characteristic at Selected Percentage Predictions of 5-Year Risk of Hip Fracture



AUC indicates area under the curve. Blue curves in Figure 1 and Figure 2 are the same and are derived from

30

Classifying Subjects Based on Predicted Values

- Graphic showing correct versus incorrect classification in test dataset (CHD prediction example)

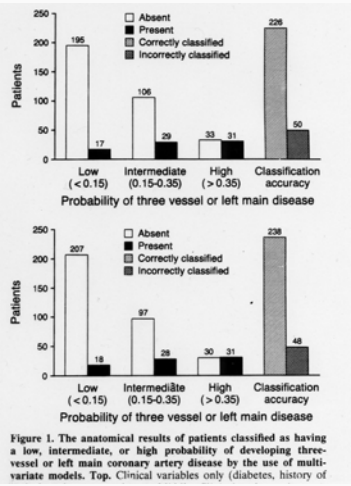


Figure 1. The anatomical results of patients classified as having a low, intermediate, or high probability of developing three-vessel or left main coronary artery disease by the use of multivariate models. Top. Clinical variables only (diabetes, history of

31

Classifying Subjects Based on Predicted Values

- Classification is really the goal of prediction models, in terms of making them clinically useful
- Classification is difficult, and is a function of underlying risk of the outcome being predicted: even relatively highly predictive models (High AUC from ROC curve, for example) may not classify well
- Classification needs to be assessed to understand clinical utility of models used: not done so often as it should be in published research
- Just like measures of relative prediction, classification summaries (sensitivity, total prediction error etc..) need to e validated on an independent test set of data: failure to do so will result in overoptimistic results

32