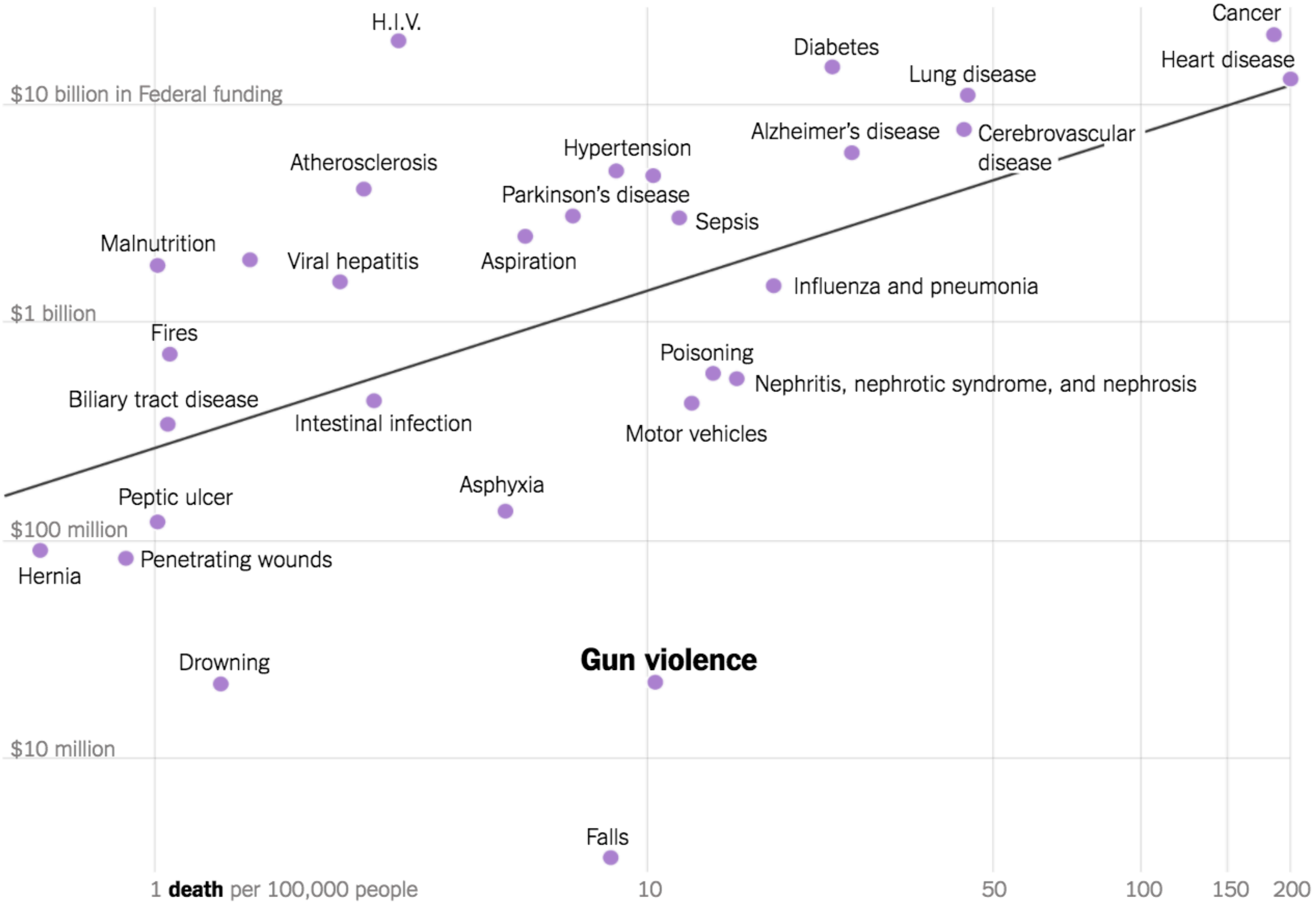
The New York Times

There’s an Awful Lot We Still Don’t Know About Guns

By **QUOCTRUNG BUI** and **MARGOT SANGER-KATZ** MARCH 2, 2018

Federal funding for research on leading causes of death



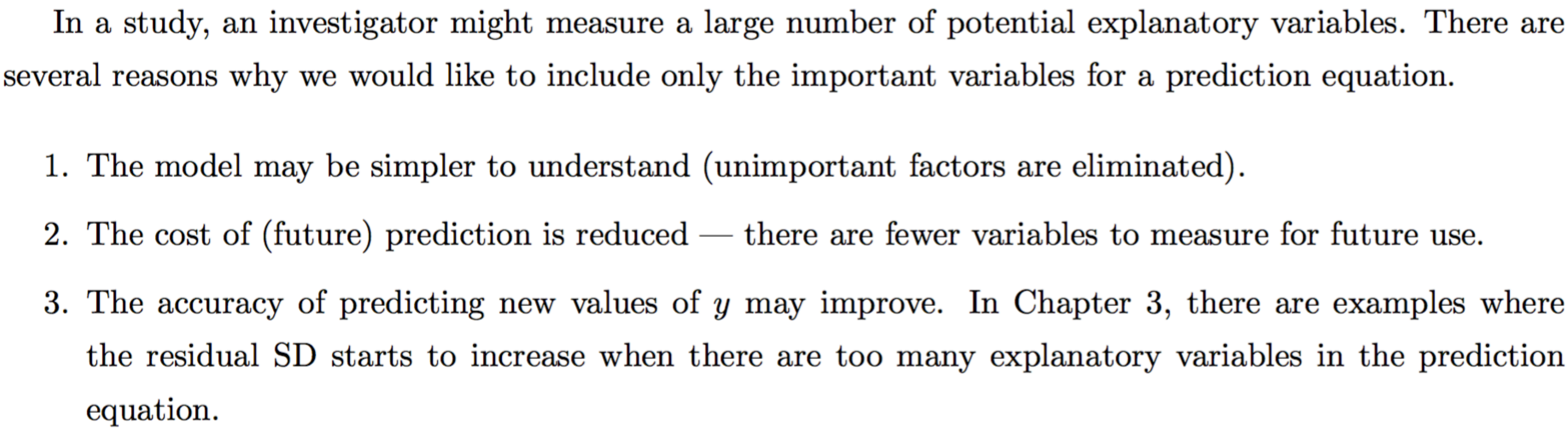
Source: From 2004 to 2014, David Stark and Nigam Shah, Funding and Publication of Research on Gun Violence and Other Leading Causes of Death

Stat 306:
Finding Relationships in Data.

Lecture 15

Sections 4.1 and 4.2

Chapter 4 – Variable selection and additional diagnostics



In a study, an investigator might measure a large number of potential explanatory variables. There are several reasons why we would like to include only the important variables for a prediction equation.

1. The model may be simpler to understand (unimportant factors are eliminated).
2. The cost of (future) prediction is reduced — there are fewer variables to measure for future use.
3. The accuracy of predicting new values of y may improve. In Chapter 3, there are examples where the residual SD starts to increase when there are too many explanatory variables in the prediction equation.

Chapter 4 – Variable selection and additional diagnostics

4.1 Variable Selection algorithms

4.2 Cross-validation and out-of sample assessment

4.3 Additional diagnostics

4.4 Transforms and nonlinearity

4.5 Diagnostics for data collected sequentially in time

Four categories of scientific study

	Observational	Experimental
Goal is Explanation	1.	2.
Goal is Prediction	3.	4.

Goal is
Explanation

1. What questions do you want to ask ?

2. Define an appropriate model.

3. Define the hypotheses that correspond to the questions of interest.

4. Collect the data.

5. Fit the model as defined earlier.

6. Answer your questions with uncertainty quantification
   (i.e. with p-values, Confidence Intervals).

Goal is
Prediction

1. What do you want to predict?

2. Define an appropriate metric for evaluating quality of predictions (e.g. RMSE, absolute prediction error, ROC curve).

3. Collect the data.

4. Separate your data into “train” and “holdout” subsets.

5. Fit many different models to the “train” subset of the data.

6. Pick the model that is “best” (according to your chosen outcome) for making predictions on the “holdout” subset of the data.

7. Note that p-values and Confidence intervals are not valid.

Goal is

Prediction... but you also want some explanations

(warning, this is a bit outdated)

1. Collect the data.

2. Select a “model-selection” criteria (e.g. Adjusted R^2 or Cp)

3. Identify all possible regression models with all possible combinations of the predictors.

4. Identify a subset of models that are best in terms of the chosen “model-selection” criteria.

5. Evaluate and refine the models identified in Step 4 by doing residual analyses, transformations, checking model assumptions.

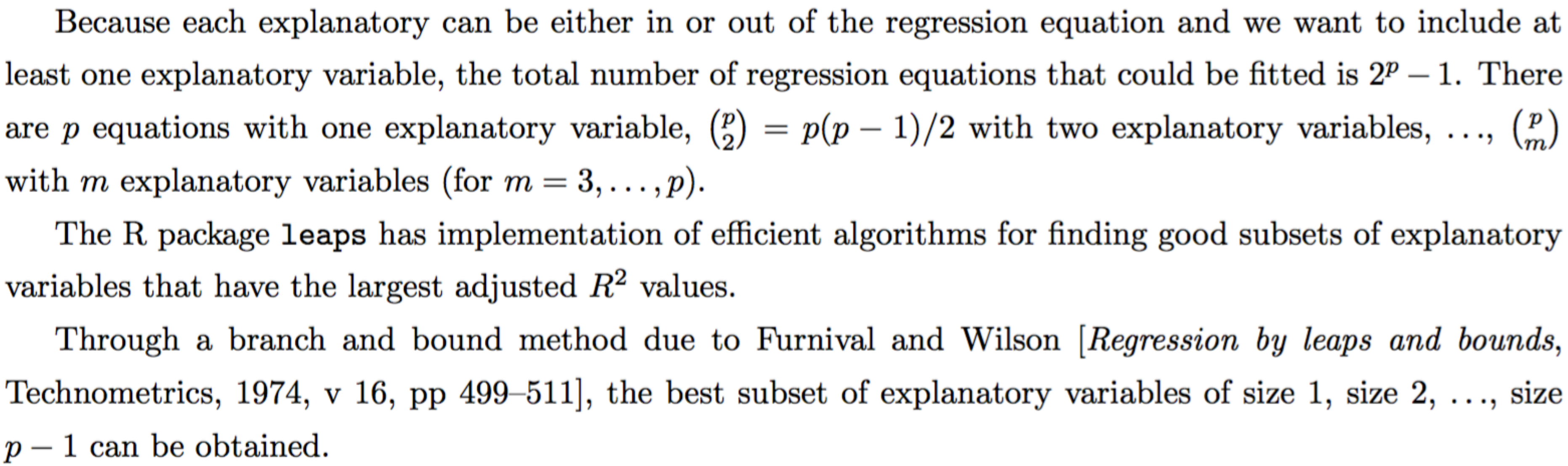
6. Pick a “best” model from the refined subset of models that meets assumptions and allows you to do some explanations.

Goal is

Prediction… but you also want some explanations

(warning, this is a bit outdated)

4.1 Variable Selection algorithms



Because each explanatory can be either in or out of the regression equation and we want to include at least one explanatory variable, the total number of regression equations that could be fitted is $2^p - 1$. There are p equations with one explanatory variable, $\binom{p}{2} = p(p-1)/2$ with two explanatory variables, …, $\binom{p}{m}$ with m explanatory variables (for $m = 3, \ldots, p$).

The R package `leaps` has implementation of efficient algorithms for finding good subsets of explanatory variables that have the largest adjusted R^2 values.

Through a branch and bound method due to Furnival and Wilson [*Regression by leaps and bounds*, Technometrics, 1974, v 16, pp 499–511], the best subset of explanatory variables of size 1, size 2, …, size $p - 1$ can be obtained.

- Even with a small number of possible covariates, there are a lot possible models one could fit.

- And think about all the possible interaction terms!

- This can make things almost impossible.

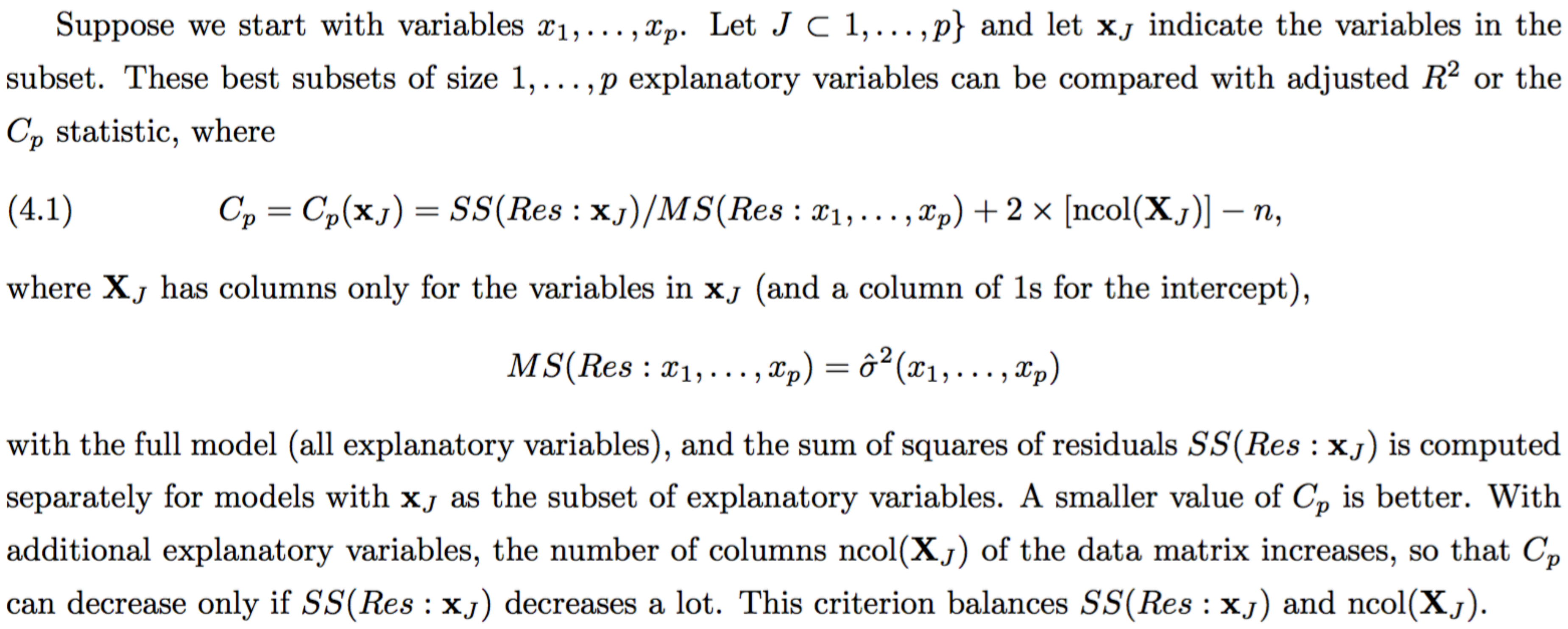
Goal is

Prediction… but you also want some explanations

(warning, this is a bit outdated)

4.1 Variable Selection algorithms

The Cp statistic, a “model-selection” criteria



Suppose we start with variables $x_1, \ldots, x_p$. Let $J \subset 1, \ldots, p\}$ and let $\mathbf{x}_J$ indicate the variables in the subset. These best subsets of size $1, \ldots, p$ explanatory variables can be compared with adjusted R^2 or the C_p statistic, where

$$C_p = C_p(\mathbf{x}_J) = SS(Res : \mathbf{x}_J)/MS(Res : x_1, \ldots, x_p) + 2 \times [\text{ncol}(\mathbf{X}_J)] - n, \tag{4.1}$$

where $\mathbf{X}_J$ has columns only for the variables in $\mathbf{x}_J$ (and a column of 1s for the intercept),

$$MS(Res : x_1, \ldots, x_p) = \hat{\sigma}^2(x_1, \ldots, x_p)$$

with the full model (all explanatory variables), and the sum of squares of residuals $SS(Res : \mathbf{x}_J)$ is computed separately for models with $\mathbf{x}_J$ as the subset of explanatory variables. A smaller value of C_p is better. With additional explanatory variables, the number of columns $\text{ncol}(\mathbf{X}_J)$ of the data matrix increases, so that C_p can decrease only if $SS(Res : \mathbf{x}_J)$ decreases a lot. This criterion balances $SS(Res : \mathbf{x}_J)$ and $\text{ncol}(\mathbf{X}_J)$.

The Cp statistic and the adjusted-R^2 are very similar

Goal is

Prediction… but you also want some explanations

(warning, this is a bit outdated)

4.1 Variable Selection algorithms

1. **Forward Selection**

2. **Backward Elimination**

Goal is

Prediction… but you also want some explanations

(warning, this is a bit outdated)

4.1 Variable Selection algorithms

1. Forward Selection

-start with one variable, add one variable at a time

2. Backward Elimination

-start with full model (all potential variables), remove one variable at a time

Goal is **Prediction**

4.2 Train / Test

Training/holdout.

Let $\hat{\boldsymbol{\beta}}^{(train)}$ be the least square regression coefficient vector for the training set of size n_1. For each row $\mathbf{x}_i^T$ of $\mathbf{X}^{(holdout)}$ in the holdout set of size n_2, get

$$\hat{y}_{i|train} = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{(train)}, \quad i \in \text{indices for holdout set.}$$

The out-of-sample root mean square (prediction) error is:

$$RMSE_{holdout}(x_1, \ldots, x_p) = \sqrt{n_2^{-1} \sum_{i \in \text{holdout}} (y_i - \hat{y}_{i|train})^2}\,.$$

$RMSE_{holdout}(\mathbf{x}_J)$ can be defined for different subsets of the explanatory variables. $[\mathbf{x}_J = (x_j : j \in J)$= variables indexed by set $J]$

How to decide on n_1, n_2. Usually $n_1 \geq n_2$.

Goal is
Prediction

1. What do you want to predict?

2. Define an appropriate metric for evaluating quality of predictions (e.g. RMSE, absolute prediction error, ROC curve).

3. Collect the data.

4. Separate your data into “train” and “holdout” subsets.

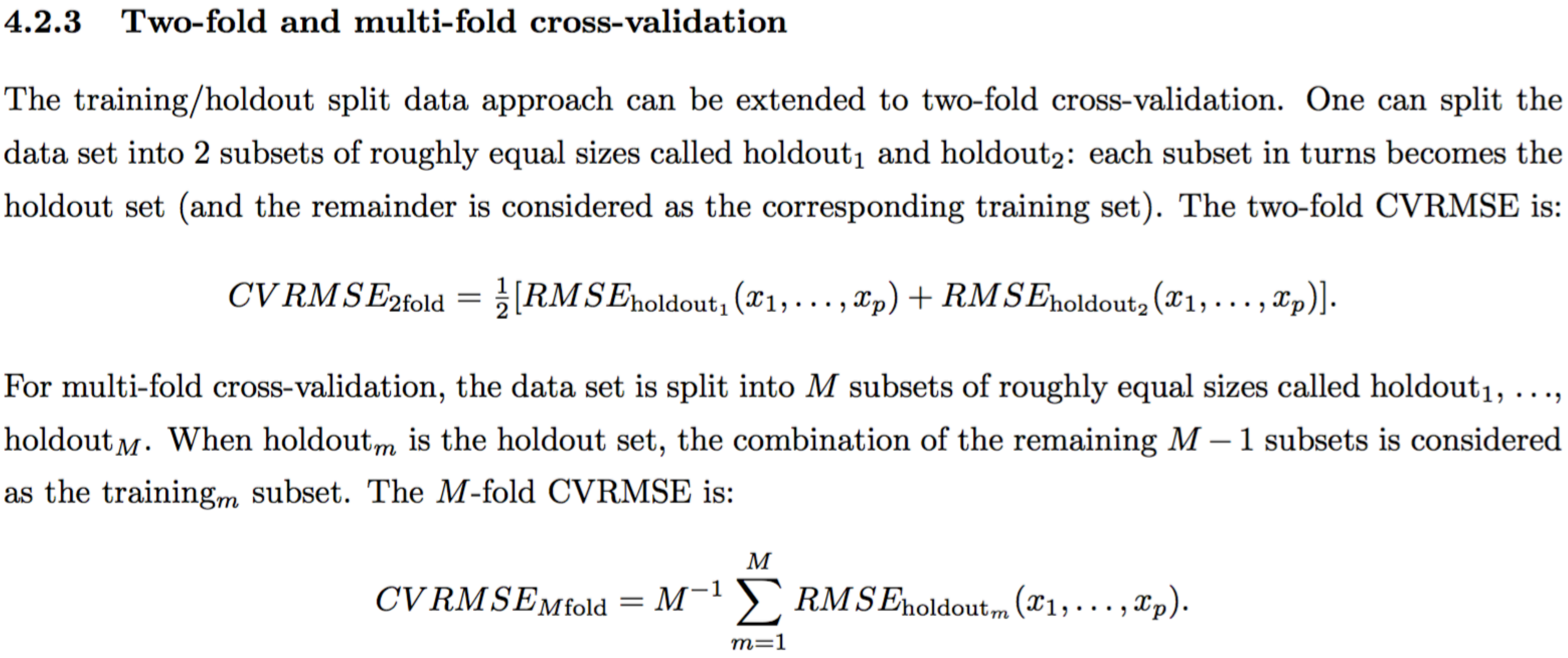
5. Fit many different models to the “train” subset of the data.

6. Pick the model that is “best” (according to your chosen outcome) for making predictions on the “holdout” subset of the data.

7. Note that p-values and Confidence intervals are not valid.

Goal is
Prediction

4.2 Cross-validation



4.2.3 Two-fold and multi-fold cross-validation

The training/holdout split data approach can be extended to two-fold cross-validation. One can split the data set into 2 subsets of roughly equal sizes called holdout_1 and holdout_2: each subset in turns becomes the holdout set (and the remainder is considered as the corresponding training set). The two-fold CVRMSE is:

$$CVRMSE_{2\text{fold}} = \tfrac{1}{2}[RMSE_{\text{holdout}_1}(x_1, \ldots, x_p) + RMSE_{\text{holdout}_2}(x_1, \ldots, x_p)].$$

For multi-fold cross-validation, the data set is split into M subsets of roughly equal sizes called $\text{holdout}_1, \ldots,$ holdout_M. When holdout_m is the holdout set, the combination of the remaining $M-1$ subsets is considered as the training_m subset. The M-fold CVRMSE is:

$$CVRMSE_{M\text{fold}} = M^{-1} \sum_{m=1}^{M} RMSE_{\text{holdout}_m}(x_1, \ldots, x_p).$$

Goal is Prediction

1. What do you want to predict?

2. Define an appropriate metric for evaluating quality of predictions (e.g. RMSE, absolute prediction error, ROC curve).

3. Collect the data.

4. Separate your data into K random subsets.

5. For k in 1:K
   - Fit your model using all the data except the kth subset.

   - Calculate metric (e.g. prediction error) based on fitting the model to the kth subset of the data.

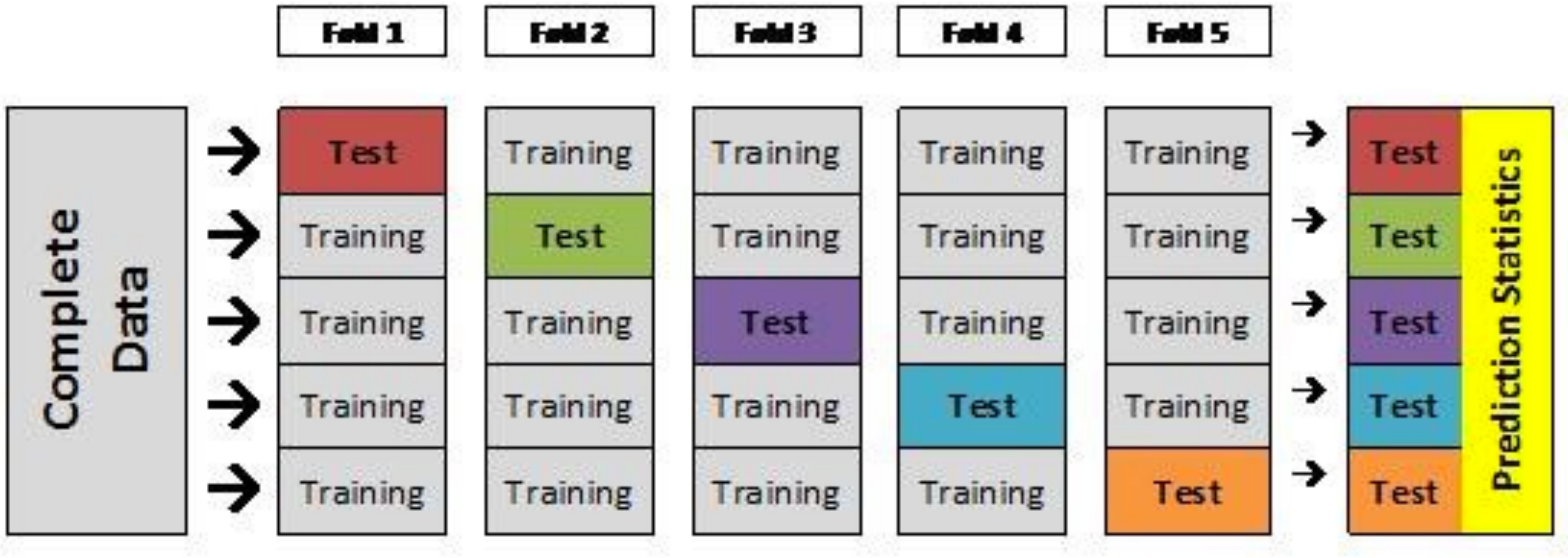
6. Calculate average of K metrics for each model.

7. Choose "best model" based on averaged metric.
8. Note that p-values and Confidence intervals are not valid.

For each model, we do 5-fold CV:

Metric:

Mean Absolute Prediction Error:



12

8

6

9

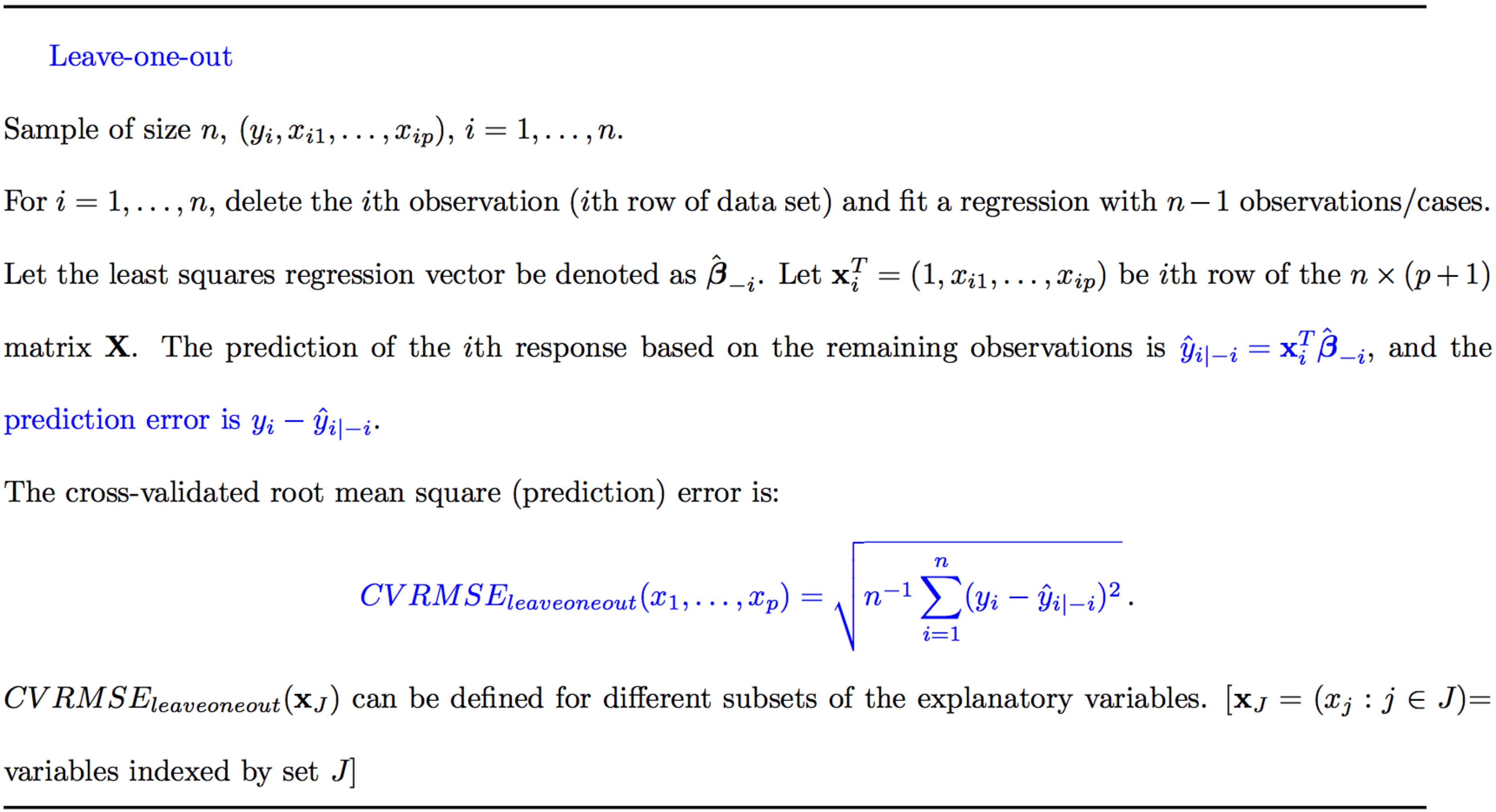
5

K-averaged metric = 40/5 = 8

Source: http://blog.goldenhelix.com/goldenadmin/cross-validation-for-genomic-prediction-in-svs/

Goal is
Prediction

4.2 Leave-one-out



Leave-one-out

Sample of size n, $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$.

For $i = 1, \ldots, n$, delete the ith observation (ith row of data set) and fit a regression with $n-1$ observations/cases.

Let the least squares regression vector be denoted as $\hat{\boldsymbol{\beta}}_{-i}$. Let $\mathbf{x}_i^T = (1, x_{i1}, \ldots, x_{ip})$ be ith row of the $n \times (p+1)$ matrix $\mathbf{X}$. The prediction of the ith response based on the remaining observations is $\hat{y}_{i|-i} = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{-i}$, and the prediction error is $y_i - \hat{y}_{i|-i}$.

The cross-validated root mean square (prediction) error is:

$$CVRMSE_{leaveoneout}(x_1, \ldots, x_p) = \sqrt{n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_{i|-i})^2}\,.$$

$CVRMSE_{leaveoneout}(\mathbf{x}_J)$ can be defined for different subsets of the explanatory variables. [$\mathbf{x}_J = (x_j : j \in J)$= variables indexed by set J]