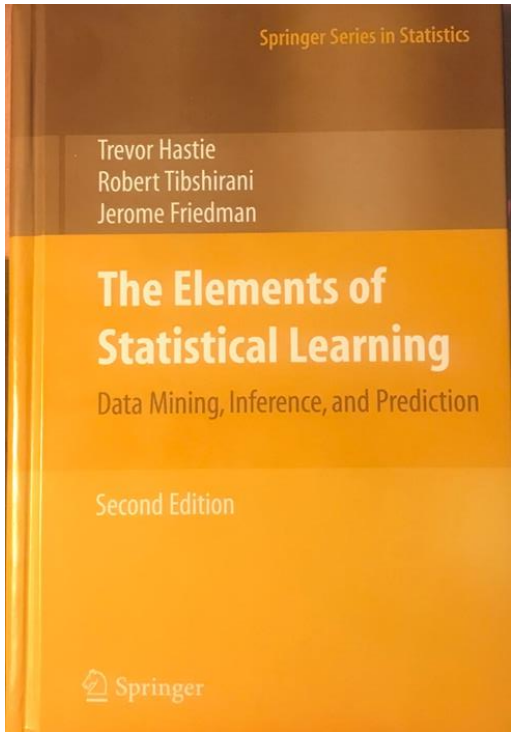
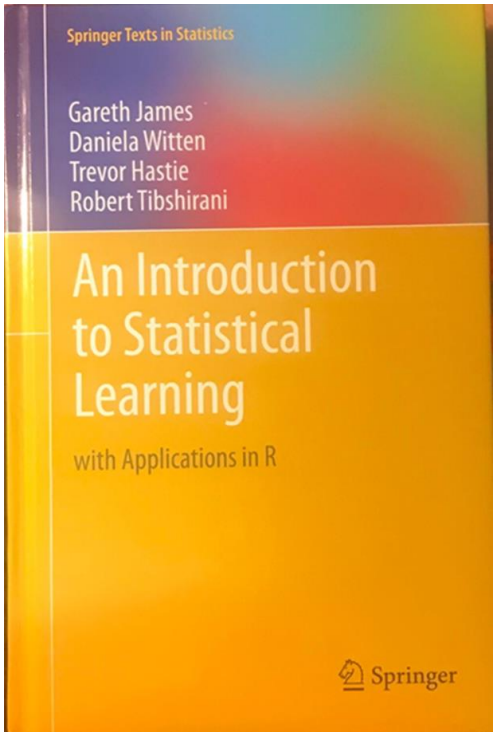
Statistical Learning for Complex Survey Data:

Using Cross-Validation for Variable Selection in Generalized Linear Models

Darryl V. Creel





I want to develop a model. How do I determine which independent variables I should include in my model?

- P-value based approaches
  - Forwards selection
  - Backwards elimination
  - Stepwise
  - Hosmer-Lemeshow
- Relative quality statistics (indirect estimation)
  - Akaike information criterion
  - Bayesian information criterion
  - Mallows Cp
- Direct estimation of the model error
  - Validation data set
  - *V*-fold cross-validation

Reminder, what is *v*-fold cross-validation? Here is an example of 5-fold cross-validation.

Iteration	Fold = 1	Fold = 2	Fold = 3	Fold = 4	Fold = 5
1	Test	Training	Training	Training	Training
2	Training	Test	Training	Training	Training
3	Training	Training	Test	Training	Training
4	Training	Training	Training	Test	Training
5	Training	Training	Training	Training	Test

How was the data generated? Population size = 10,000 and sample size = 400 using Poisson sampling

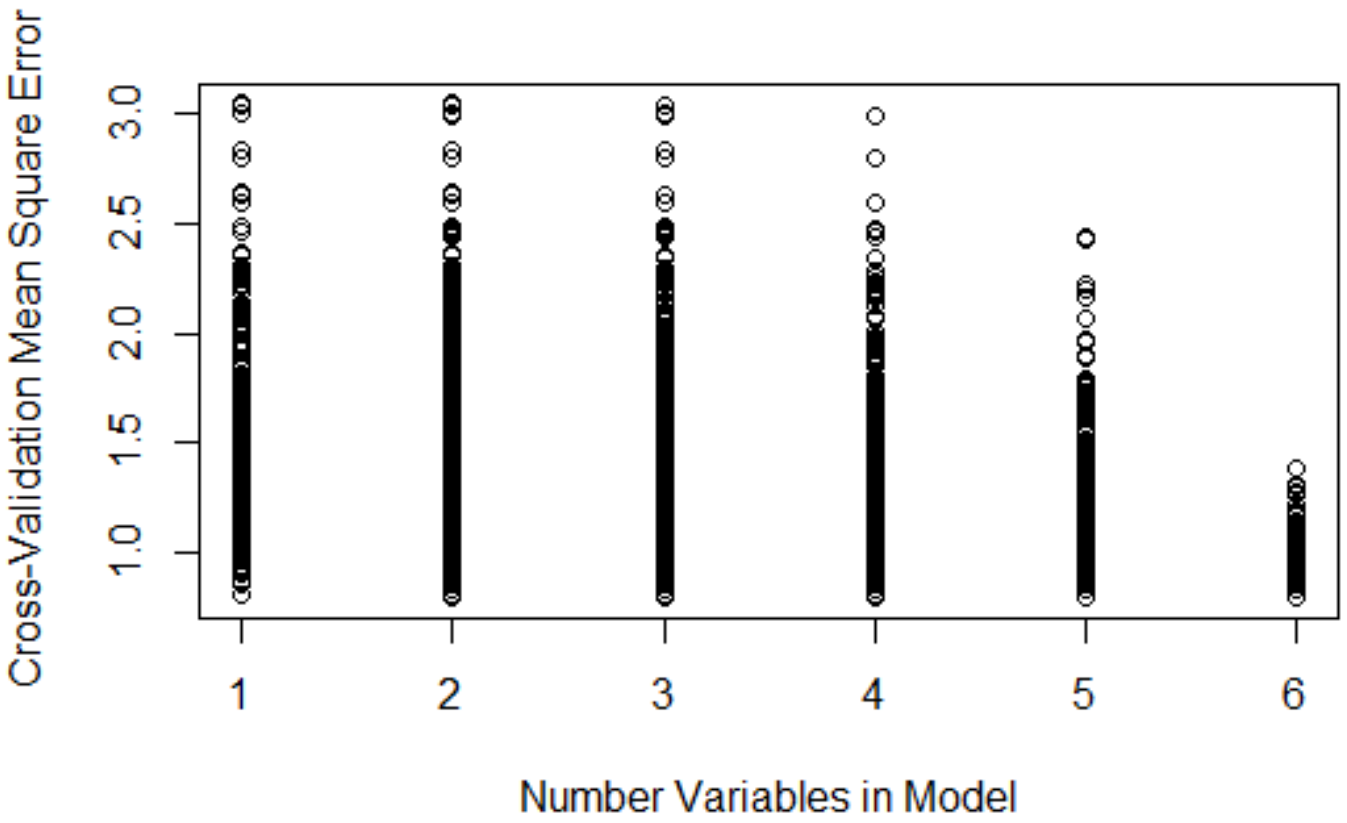
- Seven independent N(0,1) random variables, x1-x6 and error
- Three coefficients from an N(0,0.16), b1-b3
- PSI makes the error homoscedastic or hetroscedastic, 0, 0.2, 0.5

- Y = 1 + b1*x1 + b2*x2 + b3*x3 + (1 + psi*x1 + psi*x2)*error

- Informative sampling POS depends on Y
  - z <- N(1+y, 0.25)
  - k <- 1/(1+exp(2.5-0.5*z)), size variable
  - sumPopK <- sum(k)
  - probSel <- sampSize*k/sumPopK

- ranUni <- runif(n = popSize, min = 0, max = 1)
- sampInd <- ifelse(ranUni <= probSel, 1, 0)
- psuWt <- ifelse(sampInd == 1, 1/probSel, 0)

What does the distribution of cross-validation mean square error look like treating the data as if it was from a simple random sample?



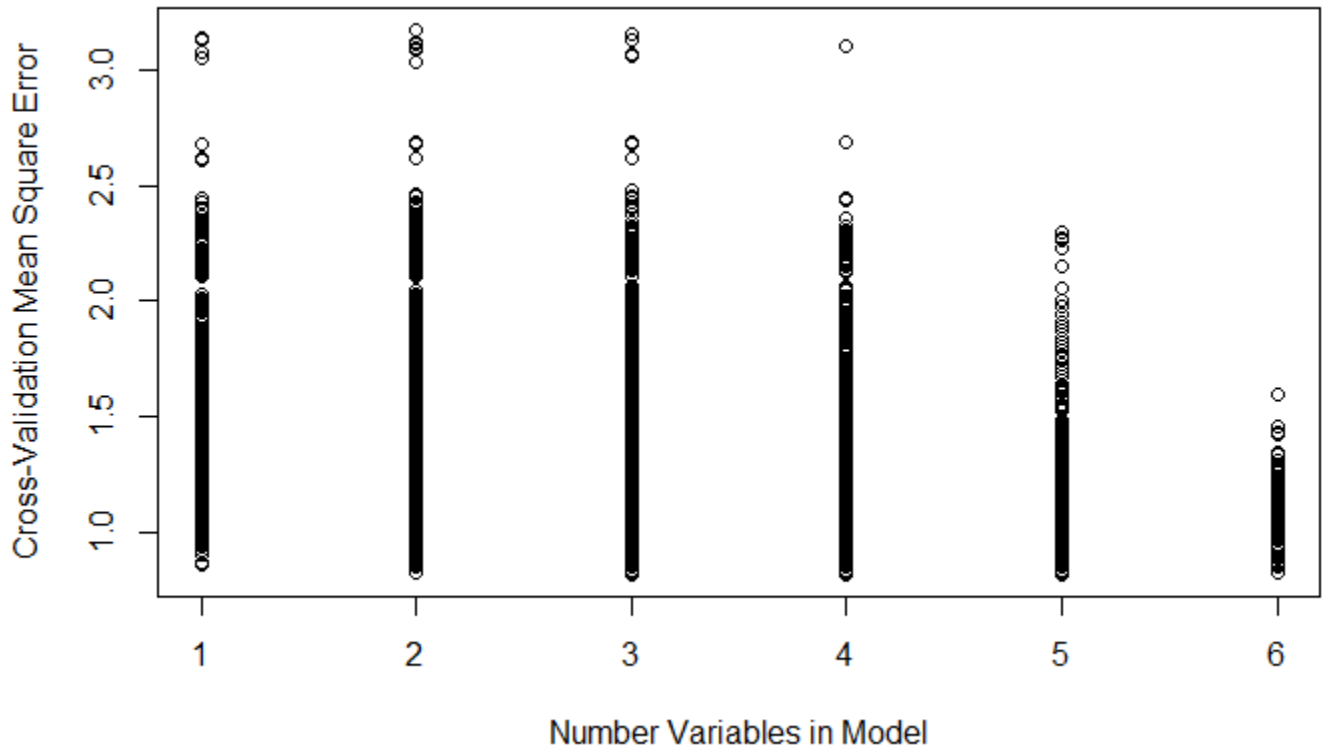
What is the minimum mean cross-validation error treating the data as if it was from a simple random sample?

numVars	cvErrorMin
<dbl>	<dbl>
1	1.33858
2	1.18070
3	1.04503
4	1.04180
5	1.03917
6	1.03687

What are the first ten models with the lowest model mean cross-validation mean square error treating the data as if it was from a simple random sample?

ModelVars	numVars	cvErrorMean
<chr>	<dbl>	<dbl>
x1 + x2 + x3 + x4 + x5 + x6	6	1.03687
x1 + x2 + x3 + x4 + x5	5	1.03917
x1 + x2 + x3 + x4 + x6	5	1.03949
x1 + x2 + x3 + x5 + x6	5	1.04010
x1 + x2 + x3 + x4	4	1.04180
x1 + x2 + x3 + x5	4	1.04238
x1 + x2 + x3 + x6	4	1.04274
x1 + x2 + x3	3	1.04503
x1 + x2 + x4 + x5 + x6	5	1.17146
x1 + x2 + x4 + x5	4	1.17421

What does the distribution of cross-validation mean square error look like for Poisson sample?



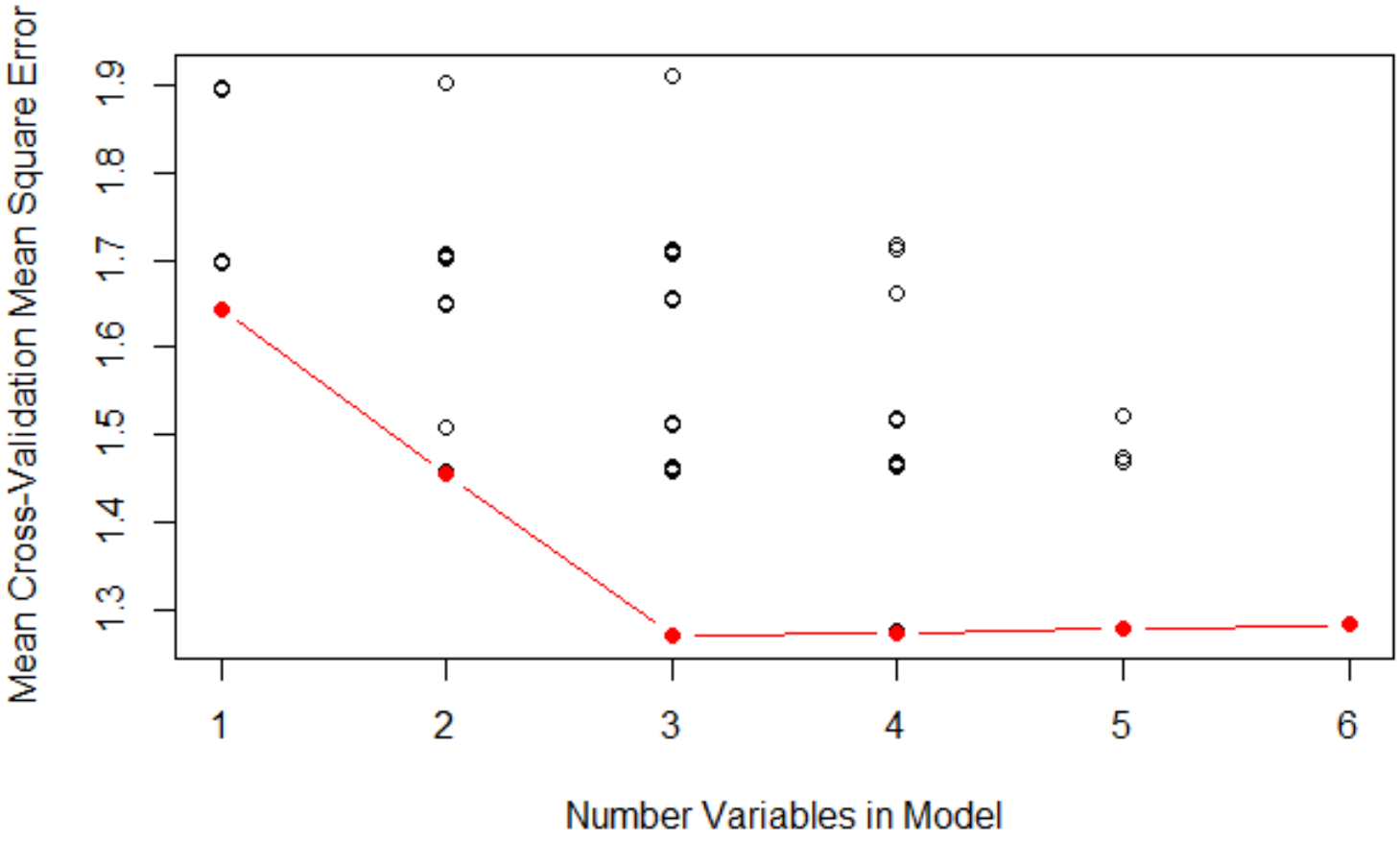
What is the minimum mean cross-validation error for a Poisson sample?

numVars	cvErrorMin
<dbl>	<dbl>
1	1.37361
2	1.22905
3	1.08761
4	1.09346
5	1.09955
6	1.10648

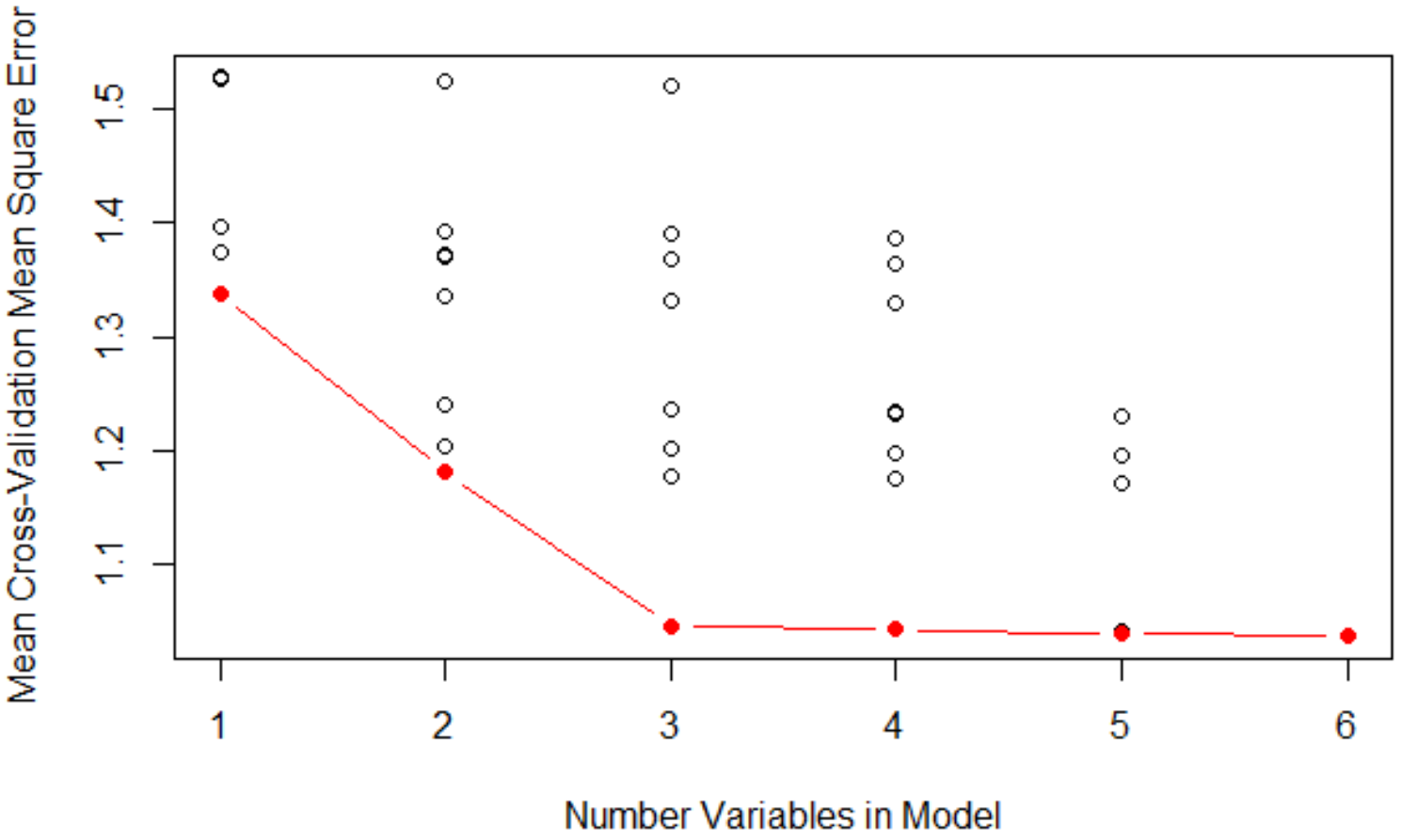
What are the first ten models with the lowest model mean cross-validation mean square error for a Poisson sample?

ModelVars	numVars	cvErrorMean
<chr>	<dbl>	<dbl>
x1 + x2 + x3	3	1.08761
x1 + x2 + x3 + x6	4	1.09346
x1 + x2 + x3 + x5	4	1.09380
x1 + x2 + x3 + x4	4	1.09444
x1 + x2 + x3 + x5 + x6	5	1.09955
x1 + x2 + x3 + x4 + x6	5	1.10015
x1 + x2 + x3 + x4 + x5	5	1.10086
x1 + x2 + x3 + x4 + x5 + x6	6	1.10648
x1 + x3	2	1.22905
x1 + x2	2	1.23545

What does distribution of the model mean of the cross-validation mean square error look like for a Poisson sample?



What does distribution of the model mean of the cross-validation mean square error look like treating the data as if it was from a simple random sample?



Considerations for v-fold cross-validation with data from a complex survey design.

- Complex survey design

- How do you create the v-folds for cross-validation?
  - Random
  - Sorted weights

- Once you have the v-folds and you start partitioning the data into training and test data based on the v-folds, how do you treat the weights?
  - Do you ignore them?
  - Use them as is?
  - Ratio adjust them to sum to the population?

  Weighted MSE

  Do not blindly apply these methods.

delivering **the promise of science**
for global good

RTI INTERNATIONAL

Name
Email
Phone Number