36-350: Data Mining

Handout 19
October 29, 2003

Assessing the quality of a regression model

How do we know how good our regression models really are?

Holdout method—Randomly divide data into "training" and "test." Lock the test data away, and do your modeling and visualization on the training data. When done, evaluate it on the test data. This provides an estimate of how well your model will perform on future data.

Problems with the holdout method:

- The model always worse than one trained on the full dataset.
- The results can depend a lot on the random split.

To fix the first problem: Instead of evaluating a particular regression model, with particular coefficients, we should focus on comparing one model type to another. For example, if I fit the linear model `Sold.5 ~ Price.5`, how well do I expect to perform versus `Sold.5 ~ Price.1 + Price.5`? (On the training set, we know the latter will win, but that is not necessarily true for future data.) This fixes the problem because we only care about the relative performance of the two types.

To fix the second problem: Average the results over multiple splits.

Cross-validation:

- Divide the data into k blocks of equal size
- For each block, fit a model to everything except that block, and test it on the block.
- After training and testing k models, average their performance to get performance of the scheme.

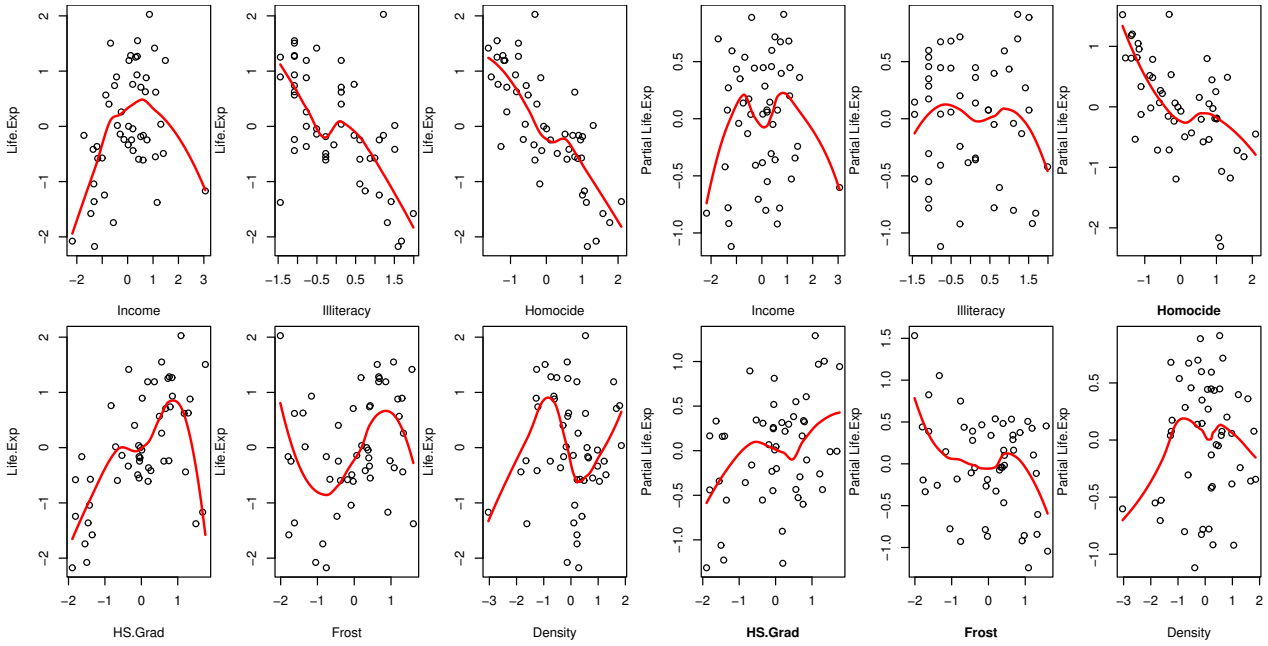
If $k = n$, then it is called **leave-one-out cross-validation**.

Results on the grocery data from handout 18:

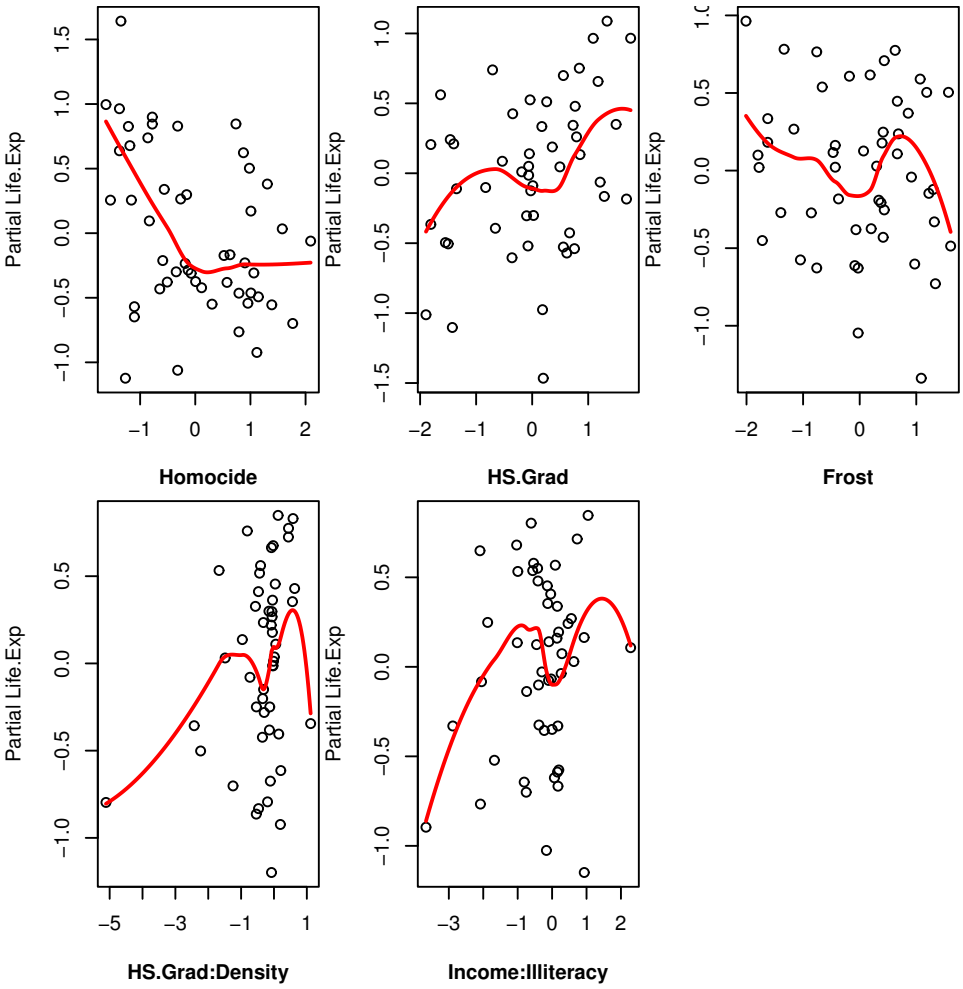
Regression model	R^2	Leave-one-out R^2	AIC
1,2,3,4,5	0.63	0.59	-104
2,3,4,5	0.63	0.59	-106
All cross terms	0.70	0.61	-110
5,min34	0.64	0.62	-112
2,5,2:3,min34	0.67	0.64	-119

Cross-validation shows the value of deleting unimportant terms, and agrees with AIC. Notice that the models with the most parameters have the largest discrepancy between R^2 and leave-one-out R^2. For comparison to the linear models, a regression tree scores 0.50, confirming what we expect for a dataset with small effects.

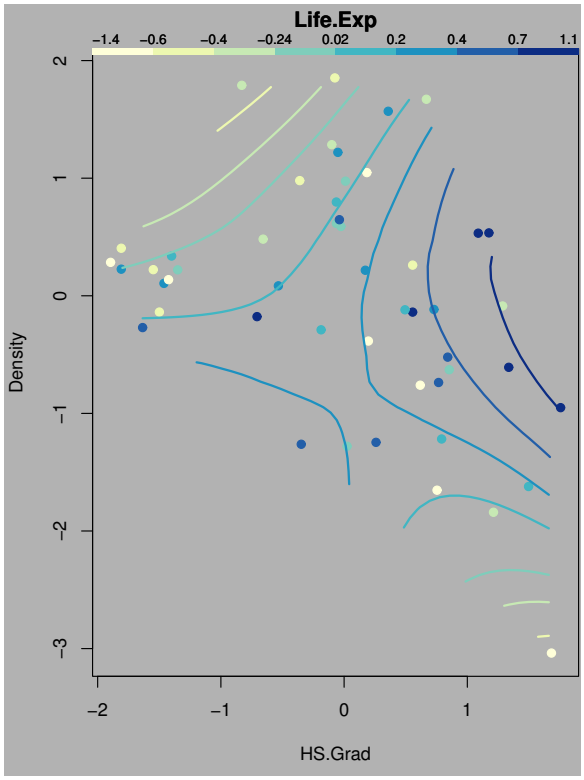
Predicting life expectancy in the 50 states:

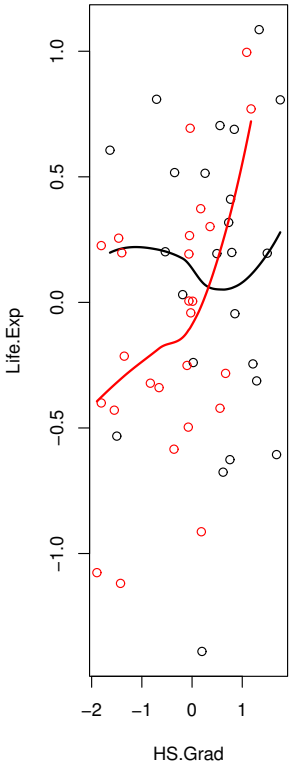


```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.14505    0.09113   1.592  0.11861
Homocide          -0.58925    0.11684  -5.043 8.35e-06 ***
HS.Grad            0.34845    0.10204   3.415  0.00138 **
Frost             -0.15028    0.10072  -1.492  0.14284
HS.Grad:Density    0.26452    0.10175   2.600  0.01265 *
Income:Illiteracy  0.13693    0.09287   1.474  0.14747
```



Regression model	R^2	Leave-one-out R^2	AIC
All predictors	0.72	0.59	-50
3 predictors	0.71	0.65	-55
3 predictors, 2 cross terms	0.76	0.69	-60


Life.Exp
−1.4 −0.6 −0.4 −0.24 0.02 0.2 0.4 0.7 1.1
Density
HS.Grad
−2 −1 0 1
2 1 0 −1 −2 −3


Life.Exp
HS.Grad
−2 −1 0 1
1.0 0.5 0.0 −0.5 −1.0


0.1−2
−3.04−0.1
Density

Describe the interaction: