

Day 4: Resampling Methods

Lucas Leemann

Essex Summer School

Introduction to Statistical Learning

- ① Motivation
- ② Cross-Validation
 - Validation Set Approach
 - LOOCV
 - k -fold Validation
- ③ Bootstrap

Resampling Methods

- Whenever we have a dataset we can sample subsets thereof - this is what *re*-sampling is. This allows us to rely in a systematic way on training and test datasets.
 - Allows to get a better estimate of the true error
 - Allows to pick the optimal model
- Sampling is computationally taxing but nowadays of little concern - nevertheless, time may be a factor.
- We will look today specifically at two approaches:
 - Cross-validation
 - Bootstrap

Validation Set Approach

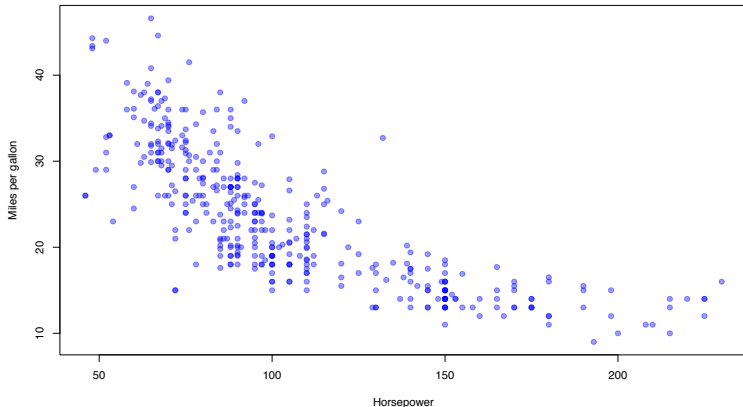
- You want to know the *true* error of a model.
- We can sample from the original dataset and create *training* and *test* dataset.
- You split the data into a *training* and a *test* dataset - you pick the optimal model on the *training* dataset and determine its performance on the *test* dataset.



(James et al, 2013: 177)

Auto Example (James et al, chapter 3)

- Predict mpg with horsepower. Problem: How complex is the relationship?



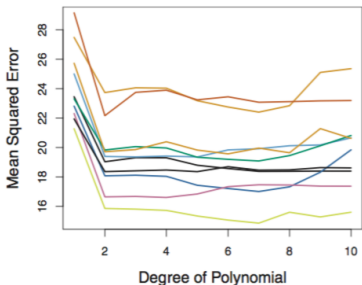
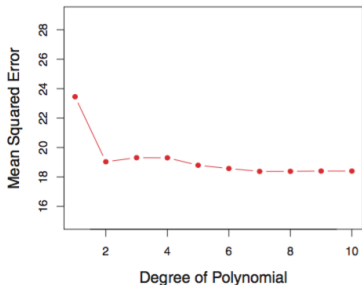
Auto Example (James et al., chapter 3) II

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
(Intercept)	39.94 *** (0.72)	56.90 *** (1.80)	60.68 *** (4.56)	47.57 *** (11.96)	-32.23 (28.57)	-162.14 * (71.43)	-489.06 * (189.83)
horsepower	-0.16 *** (0.01)	-0.47 *** (0.03)	-0.57 *** (0.12)	-0.08 (0.43)	3.70 ** (1.30)	11.24 ** (4.02)	33.25 ** (12.51)
horsepower2		0.00 *** (0.00)	0.00 * (0.00)	-0.00 (0.01)	-0.07 ** (0.02)	-0.24 ** (0.09)	-0.85 * (0.34)
horsepower3			-0.00 (0.00)	0.00 (0.00)	0.00 ** (0.00)	0.00 * (0.00)	0.01 * (0.00)
horsepower4				-0.00 (0.00)	-0.00 ** (0.00)	-0.00 * (0.00)	-0.00 * (0.00)
horsepower5					0.00 ** (0.00)	0.00 * (0.00)	0.00 * (0.00)
horsepower6						-0.00 * (0.00)	-0.00 (0.00)
horsepower7							0.00 (0.00)
R ²	0.61	0.69	0.69	0.69	0.70	0.70	0.70
RMSE	4.91	4.37	4.37	4.37	4.33	4.31	4.30

*** p < 0.001, ** p < 0.01, * p < 0.05

How many polynomials should be included?

Validation approach applied to Auto



(James et al, 2013: 178)

- Validation approach: highly variable results (right plot)
- Validation approach may tend to over-estimate test error due to small sample for training data.

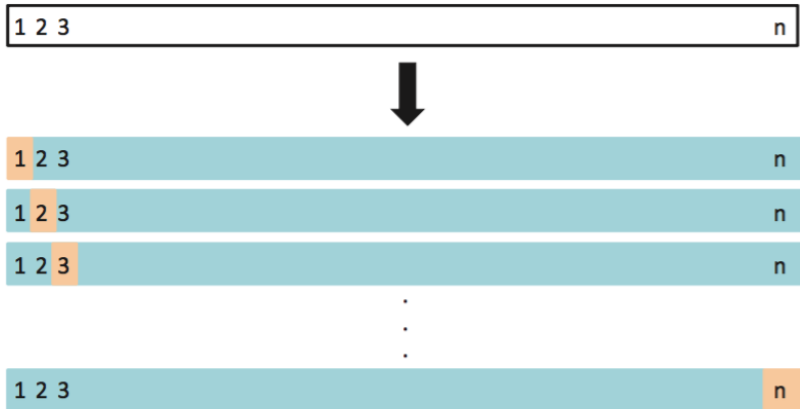
LOOCV 1

- Disadvantage 1: The error rate is highly variable
- Disadvantage 2: A large part of the data are not used to train the model

Alternative approach: Leave-one-out-cross-validation

- Leave on out and estimate model, assess the error rate (MSE_i)
- Average over all n steps, $CV_n = \frac{1}{n} \sum_{i=1}^n MSE_i$

LOOCV 2



(James et al, 2013: 179)

LOOCV 3

For LS linear or polynomial models there is a shortcut for LOOCV:

$$CV_{LOOCV} = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - h_i} \right)^2$$

Advantages:

- Less bias than validation set approach - will not over-estimate the test error.
- The *MSE* of LOOCV does not vary over several attempts.

Disadvantage:

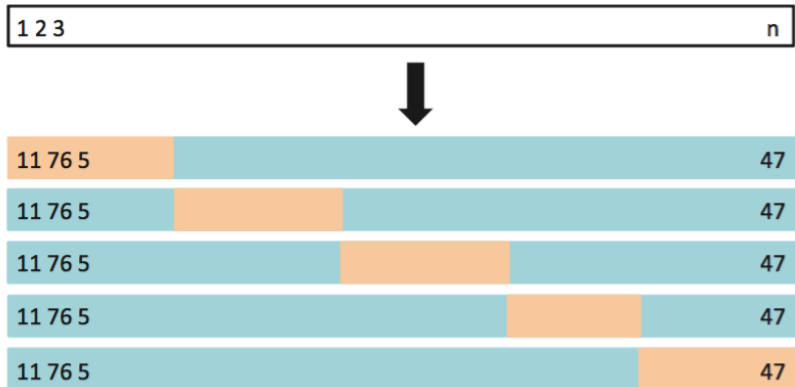
- One has to estimate the model n times.

k -fold Validation

- Compromise between validation set and LOOCV is k -fold validation.
- We divide the dataset into k different folds, whereas $k = 5$ or $k = 10$.
- We then estimate the model on $d - 1$ folds and use the k th fold as test dataset:

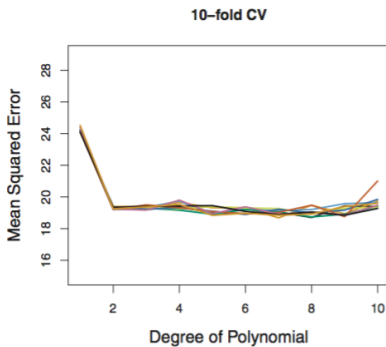
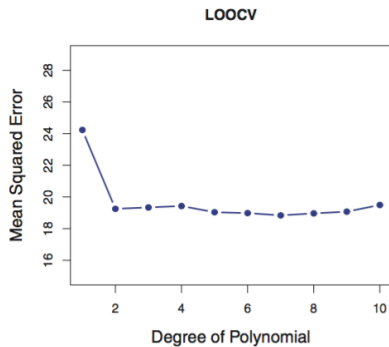
$$CV_k = \frac{1}{k} \sum_{i=1}^K MSE_i$$

k -fold validation



(James et al, 2013: 181)

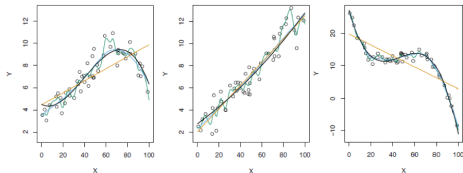
k -fold validation vs LOOCV



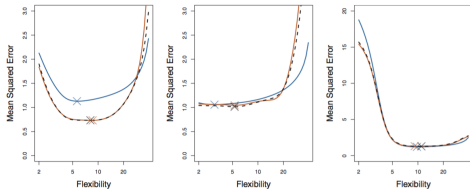
(James et al, 2013: 180)

Note: Similar error rates, but 10-fold CV is much faster.

k -fold validation vs LOOCV



(James et al, 2013: ch2)



blue: true MSE
black: LOOCV MSE
brown: 10-fold CV

(James et al, 2013: 182)

Variance-Bias Trade-Off

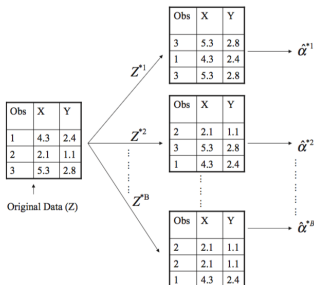
- LOOCV and k -fold CV lead to **estimates** of the test error.
- LOOCV has almost no bias, k -fold CV has small bias (since not $n - 1$ but only $(k - 1)/k \cdot n$ observations used for estimation).
- But, LOOCV has higher variance since all n data subsets are highly similar and hence the estimates are stronger correlated than for k -fold CV.
- Variance-Bias trade-off: We often rely on k -form for $k = 5$ or $k = 10$.

CV Above All Else?

- CV is fantastic but not a silver bullet.
- It has been shown that CV does not necessarily work well for hierarchical data:
 - One problem is to create independent folds (see Chu and Marron, 1991 and Alfons, 2012)
 - CV not well suited for model comparison of hierarchical models (Wang and Gelman, 2014)
- One alternative: Ensemble Bayesian Model Averaging (Montgomery et al., 2015 and see for MLM Broniecki et al., 2017).

Bootstrap

- Bootstrap allows us to assess the certainty/uncertainty of our estimates with one sample.
- For standard quantities like $\hat{\beta}$ we know how to compute $se(\hat{\beta})$. What about other non-standard quantities?
- We can re-sample from the original samples:



(James et al, 2013: 190)

Bootstrap (2)

```
> m1 <- lm(mpg ~ year, data=Auto)
> summary(m1)
```

Residuals:

Min	1Q	Median	3Q	Max
-12.0212	-5.4411	-0.4412	4.9739	18.2088

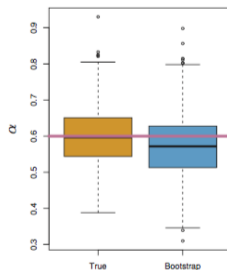
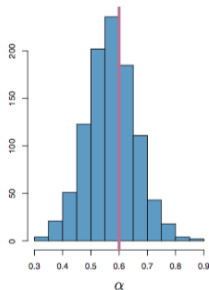
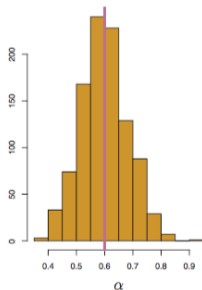
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-70.01167	6.64516	-10.54	<2e-16 ***
year	1.23004	0.08736	14.08	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> set.seed(112)
> n.sim <- 10000
> beta.catcher <- matrix(NA,n.sim,2)
> for (i in 1:n.sim){
+   rows.d1 <- sample(c(1:392),392,replace = TRUE)
+   d1 <- Auto[rows.d1,]
+   beta.catcher[i,] <- coef(lm(mpg ~ year, data=d1))
+ }
>
> sqrt(var(beta.catcher[,1]))
[1] 6.429225
```

Bootstrap (3)



yellow: 1,000 datasets
blue: 1,000 bootstrap samples

(James et al, 2013: 189)

Lab

- Cross-validation (LOOCV, and k-fold)
- Bootstrap
- CV applied to classification