Part 7: Model selection, regularization and dimension reduction

Source: James, Witten, Hastie, Tibshirani (2013)
An Introduction to Statistical Learning (ISLR)
Ch 6 + 10.2
(Bingham & Fry (2010) Ch. 4.6, 7.1)

The MLR-model: $Y = X\beta + \varepsilon$, $\varepsilon \sim N_n(0, \sigma^2 I)$

The MLR-model is very simpel, but easy to interpret.
It is often shown to give good predictive performance.
↑ to use the estimated MLR model on new data.

Up to now: we have used least squares (LS) to fit a given data set of covariates.

Now: "replace" LS fitting with alternative procedures to achieve better interpretability (feature selection), and better prediction accuracy (expecially when $p > n$)
↖ new fields of research genomics

Three main topics:

1) Subset selection ← best subset, stepwise procedures [ISLR 6.1, BF 7.1]
   ↑ covariates

2) Shrinkage (regularization): increase bias to reduce variability [ISLR 6.2]
   ← lasso, ridge

3) Dimension reduction: study linear combinations of predictors
   ← principal component analysis, partial least squares [ISLR 10.2, 6.3]

Note on notation

a) BF: β_1 intercept, p parameters including intercept

DOE, ISLR: β_0 intercept, $(p+1) - 1$

b) BF, DOE: $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$

ISLR: $RSS = SSE$

and as a consequence of a) $MSE = \frac{SSE}{n - \#\text{param}}$

$\#\text{param}$: p or $(p+1)$

Subset selection [ISLR 6.1]

best subsets
stepwise

Best subset selection algorithm

1) M_0 : model where $Y = \beta_0 + \varepsilon \Leftrightarrow \hat{\beta}_0 = \bar{y}$

2) For $k = 1, \dots, p$: M_k best model with k predictors.

best = smallest SSE, or largest R^2

LS : min SSE

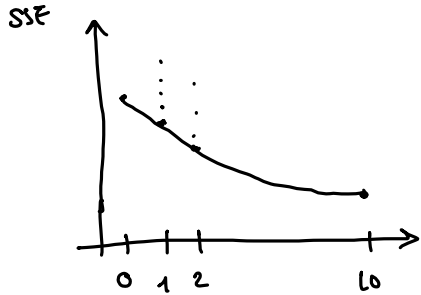
$$R^2 = 1 - \frac{SSE}{SST}$$

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$SST = \sum (y_i - \bar{y})^2$$

↑ constant over all model

min SSE $\Leftrightarrow$ max R^2



Credit -data set

SSE will always decrease (or be constant) as more covariates are added to the model, even if the new covariate is just random noise.

This is why we may not use SSE to choose between models with different number of covariates.

3) Select the best model among $M_0, M_1, \ldots, M_p$ using

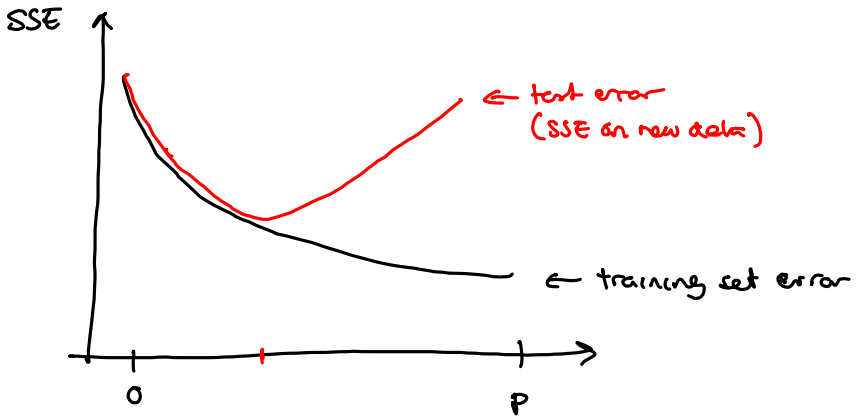
R^2_{adj}, Mallows C_p, AIC, BIC or other criteria (that take into account the number of covariates)

In ISLR much effort is put into explaining the concept of training error and test error.

The data set in hand is referred to as "training data", while future new data (for prediction) is refered to as test data.

Training error: SSE on our data set

Test error: SSE on a new data set, but using the parameter estimates from the training set.



Stepwise selection

→ used instead of best subsets when p is very large ($p > 30$).

$p = 10$: $2^{10} = 1024$

$p = 40$: $2^{40} = 1e12$

Forward stepwise selection

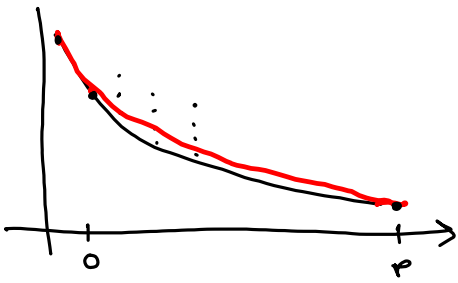
1) M_0

2) for $k = 0, \dots, p-1$

   Have best M_k, only explore adding all possible covariates one at a time

   Choose M_{k+1} with smallest SSE

3) As before.



Q: How many models are forward stepwise evaluating?

0	1	intercept
1	p	
2	p-1	
3	p-2	
	⋮	
p	1	

$\sum_{i=1}^{p} i = \frac{p(p+1)}{2}$

	best subset $2^p + 1$	$\frac{p(p+1)}{2} + 1$
$p = 10$	1025	56
$p = 50$	$1.12 \cdot 10^{15}$	1276

The principle of parsimony

"if two models are not very different, then always choose the simplest one"

C_p, AIC, BIC og R^2_{adj}

These are methods that try to mimic the behaviour of a test set – but only using the training set.

All includes a penalty on the number of peremeters fitted

Adjusted R^2

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

(β_0 and other peremeters; $\leftarrow$ df from Ch 3)

No formal reason for this choice, but easy to use and popular.

Mallow's C_p [BF Ch 7.1]

d = # pars estimated

BF: $$C_p' = \frac{SSE}{\hat{\sigma}^2} - (n - 2d)$$

$\hat{\sigma}^2$ is an estimate of σ^2 obtained from a subjective choice of a full model.

In practise the model with all covariates ($RSE_{full} = \hat{\sigma}^2$)

$$E(C_p') \approx d$$

C_p' is then plotted against d and good models lie close to the line.

ISLR use $$C_p = \frac{\hat{\sigma}^2}{n}(C_p' + n)$$

$$C_p = \frac{1}{n}(SSE + 2d\hat{\sigma}^2)$$

Interpretation: C_p adds penalty to SSE to adjust for the fact that it underestimates the test error.