Linear model selection and regularization *

for least square → true relationship is linear → low bias
→ $n \gg p$ → low variance → otherwise high variance and overfitting:
→ $p > n$ → no unique solution

Alternative → Subset ~~solution~~ selections → select subset of p → use least square on them
→ Shrinkage (regularization) → fit on all p → coefficients go towards 0 → reduce variance
→ can also be viewed as variable selection
→ Dimension reduction → projecting p predictors on M-dimension ($M<P$) where M are linear combinations of variables. Then fit on the M predictors.

Subset selection

- Best subset selection

1. Let M_0 denote the null model wich contains no predictors. This model simply predicts the sample mean for each observations
2. For $k = 1, 2, \ldots p$
   (a) Fit all $\binom{p}{k}$ models that contain k predictors
   (b) Pick the best among $\binom{p}{k}$ models, and call it M_k: (smallest RSS or biggest R^2).
3. Select the best among $M_0, \ldots M_p$ using C_p, BIC, adjusted R^2

- Forward stepwise

1. Let M_0 denote the null models with no predictors
2. For $k = 0, \ldots p-1$
   (a) consider all $p-k$ models that augment the predictors in M_k with one additional predictors.
   (b) choose the best among these $p-k$ modes, calls it M_{k+1} (using RSS or R^2)
3. Select best among $M_0 \ldots M_p$. (BIC, AIC, ad R^2)

- Backward Stepwise

1. Let M_p denote the full model, which contains all p predictors.
2. For $k = p, p-1, \cdots 1$:
   (a) consider all k models that contain all but one of the predictors in M_k, for a total of $k-1$ pred.
   (b) choose the best among these k models and call it M_{k-1}. Here using RSS or R^2
3. choose best among $M_0, \ldots M_p$

- Select set of models with ≠ numbers of variables

best is small
- $C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$ where $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ. Used for a fitted LS.
- The AIC is defined for fit by maximum likelihood. In the case of the model with gaussian errors, maximum likelihood and LS are the same $AIC = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$.
- BIC take a bayesian point of view → $BIC = \frac{1}{n}(RSS + \log(n)\, d\, \hat{\sigma}^2)$.

best is big.
- For a least squares adjusted $R^2 = 1 - \dfrac{RSS/(n-d-1)}{TSS/(n-1)}$

Shrinkage method

Ridge regression → $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$ → tend to zero → cross validation to select λ

As $\lambda \nearrow$ the flexibility $\searrow$, leading to $\searrow$ variance but $\nearrow$ ~~on~~ bias → good to use when LS has big variance.

The lasso

$$\sum_{i=1}^{n}\Big(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\Big)^2+\lambda\sum_{j=1}^{p}|\beta_j| = RSS+\lambda\sum_{j=1}^{p}|\beta_j|$$

→ l_1 norm set some to 0 when λ is high enough. → work as variable selection and model easier to interpret.

Dimension reduction methods

- Let $Z_1, \dots Z_M$ represent $M < p$ linear combinations of our original p predictors that is $Z_m = \sum_{j=1}^{p} \phi_{jm} x_j$ and we then fit $y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i$

- PCA is a technique for reducing the dimension of an $n \times p$ data matrix X. The first principal component direction of the data is that along which the observation varies the most. component vector defines the line that is as close as possible to the data.

- The principal component regression approach involves constructing the first M principal components, $Z_1, \dots Z_M$ and then using these components as the predictors in a linear regression model that is fit using LS. The key idea is that often a small number of principal components suffice to explain most of the variability in the data, as well the relationship with the response. In other words, we assume that the directions in which $X_1, \dots X_p$ show the most variation are the directions that are associated with Y. PCR is not a selection method because each of the M principal components is a linear combination of all p of the original features → number of M chosen by cross-validation.

- In PCR the direction are identified in unsupervised way, since the response Y is not used. Hence there is no guarantee that the directions that best explain the predictors will also be the best direction for predicting the response.

- Partial least squares does like PCR but use Y to select; it reduce bias but increase variance

Considerations in high dimension

- C_p, AIC, BIC approaches are not appropriate in the high-dimensional setting because estimating $\hat{\sigma}^2$ is problematic. Same for $\overset{d}{R}^2$
- fitting less flexible model are useful for performing regression in the high dimensional setting
- In the high-dimensional setting, the multicolinearity problem is extreme: any variable in the model can be written as a linear combination of all other variables. → we can never know exactly which variables (if any) truly are predictive of the outcomes, and we can never identify the best coefficients for use in the regression. At most, we can hope to assign large regression coefficients to variables that are correlated with the variables that truly are predictive of the outcome
- when $p > n$ it is easy to obtain a useless model that has zero residuals → hence we should never use sum of squared errors, p-values, R^2 statistics in high dimension. → It is important to instead report results on an independent test-set or cross validation errors, for instance, the MSE or R^2 on an independent test set is a valid measure.