**Best Subset**

For each $k \in \{0, \ldots, p\}$ there are

$$\binom{p}{k}$$

different models with $k$ predictors excluding the intercept,and $p - k$ parameters $= 0$.

There are in total

$$\sum_{k=0}^{p} \binom{p}{k} = 2^p$$

different models. For the prostate dataset with $2^8 = 256$ possible models we can go through all models in a split second. With $2^{40} = 1.099.511.627.776$ we approach the boundary.

**Subset Selection – A Constraint Optimization Problem**

Let $L_r^k$ for $r = 1, \ldots, \binom{p}{k}$ denote all $k$-dimensional subspaces of the form

$$L_r^k = \{\beta \mid p - k \text{ coordinates in } \beta = 0\}.$$

$$\hat{\beta}^k = \underset{\beta \in \cup_r L_r^k}{\operatorname{argmin}} \operatorname{RSS}(\beta)$$

The set $\cup_r L_r^k$ is *not* convex – local optimality does not imply global optimality.

We can essentially only solve this problem by solving all the $\binom{p}{k}$ subproblems, which are convex optimization problems.

Conclusion: Subset selection scales computationally badly with the dimension $p$. *Branch-and-bound* algorithms can help a little ...

**Figure 3.5 – Best Subset Selection**

The residual sum of squares $\operatorname{RSS}(\hat{\beta}^k)$ is a monotonely decreasing function in $k$.

The selected models are in general *not nested*.

One can not use $\operatorname{RSS}(\hat{\beta}^k)$ to select the appropriate subset size only the best model of subset size $k$ for each $k$.

*Model selection criterias* such as *AIC* and *Cross-Validation* can be used – these are major topics later in the course.

**Test Based Selection**

Set

$$\hat{\beta}^{k,r} = \underset{\beta \in L_r^k}{\operatorname{argmin}} \operatorname{RSS}(\beta)$$

and fix $L_s^l \subseteq L_r^k$ with $l < k$.

$$F = \frac{(N-k)[\operatorname{RSS}(\hat{\beta}^{l,s}) - \operatorname{RSS}(\hat{\beta}^{k,r})]}{(k-l)\operatorname{RSS}(\hat{\beta}^{k,r})}$$

follows under *Assumption 2* an F-distribution with $(k-l, N-k)$ degrees of freedom *if* $\beta \in L_s^l$.

$L_r^k$ is preferred over $L_s^l$ if $\Pr(. > F) \leq 0.05$, say – the deviance from $L_s^l$ is unlikely to be explained by randomness alone.

**Test Based Selection – Pros and Cons**

- *Plusses*
  - We can control the *type I error* for two *a priori specified, nested models.*
  - We can control the *total type I error* for sequentially testing a *sequence of a priori specified, nested models.*
- *Minusses*
  - Non-nested models are in-comparable.
  - We do not understand the distributions of multiple a priori non-nested tests.
  - We don't control the power, only the type I error ...
  - ... and tests are by nature asymmetric, a complex model is accepted over a simple model when the simple model is ?clearly? *inadequate.*

*Take home message:* Test statistics are useful for quantifying if a simple model is inadequate compared to a complex, but there is *no theoretical foundation* for general test based model selection strategies.

We write ?clearly? because how clear it is depends upon the level chosen for the test. At level $\alpha = 0.05$ it is perhaps an exaggeration to say clearly. Using level 0.05 we might say that there is some but not overwhelming evidence that the simple model is inadequate. However, we typically compute a $p$-value, which we regard as a quantification of how inadequate the simple model is. In reality, many tests are then rejected with $p$-values that are very small, in which case we can say that there is clear evidence. If the $p$-value is close to 0.05, we must draw our conclusions with caution.

The formal statistical tests belong most naturally in the world of *confirmatory statistical analysis* where we want to confirm a specific, a priori given hypothesis – for instance the effect or superiority of a given drug. The theory of hypothesis testing is not really developed for model selection.

**Strategies for Approximating Best Subset Solutions**

If we can't go through all possible models we need approximations.

- *Forward stepwise selection.*
  - Initiate using only the intercept
  - In the $k$'th step include the variable among the $p-k$ remaining that improves RSS the most.
- *Backward stepwise selection.*
  - Initiate by fitting the full model – requires $N \geq p$
  - In the $k$'th step exclude the variable among the $k$ remaining that increases RSS the least.
- *Mixed stepwise strategies.*
  - For instance, initiate by fitting the full model.
  - In the $r$'th step include or exclude a variable according to a tradeoff criteria between the best improvement or least increase in RSS.

**Penalized Regression**

If $J : \mathbb{R}^p \to [0, \infty)$ is any function we replace the least squares estimate by the *penalized least squares estimate*

$$\hat{\beta}^{\lambda J} = \operatorname*{argmin}_{\beta} \mathrm{RSS}(\beta) + \lambda J(\beta).$$

The optimization problem is nicest if $J$ is convex. The parameter $\lambda \geq 0$ determines the tradeoff between the *measure of fit to data*, RSS, and the *penalty on the parameter*, $J$.

The function $J$ implements an a priori preference of some parameters over other. It is the frequentists version of a Bayesian incorporation of prior beliefs.

To a Bayesian we are computing the posterior mode when we use the prior

$$c(\lambda/2\sigma^2)^{-1} \exp\left(-\frac{\lambda}{2\sigma^2} J(\beta)\right), \quad c(\lambda) = \int \exp(-\lambda J(\beta)) \mathrm{d}\beta$$

on the mean value parameter $\beta$.

Of course, we need $c(\lambda) < \infty$ for the Bayesian interpretation – otherwise we don't have a proper prior. But in this case we see that the minimizer $\tilde{\beta}$ is the minimizer of minus-log-posterior, and since the minus-log function is monotonely decreasing $\hat{\beta}^J$ is the actually the mode of the posterior

$$\frac{1}{c(\lambda/2\sigma^2)\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left[(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda J(\beta)\right]\right).$$

**Ridge Regression**

If $J(\beta) = \beta^T\beta = ||\beta||^2$ the penalized estimation method is known as *ridge regression*.

We need to optimize

$$(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta.$$

The function $J$ is *strictly convex* with $J(\beta) \to \infty$ for $||\beta|| \to \infty$. There is *always* a unique minimum $\hat{\beta}^{\text{ridge}}$ when $\lambda > 0$.

**Ridge Regression – The Solution**

Observe that by augmenting $\mathbf{y}$ with $p$ trailing zero's and $\mathbf{X}$ with a trailing $p \times p$ matrix $\sqrt{\lambda}I_p$ we get

$$\left(\begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}I_p \end{bmatrix}\beta\right)^T \left(\begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}I_p \end{bmatrix}\beta\right) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta.$$

The minimization is an ordinary least squares problem with solution

$$\begin{aligned} \hat{\beta}^{\text{ridge}} &= \left(\begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}I_p \end{bmatrix}^T \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}I_p \end{bmatrix}\right)^{-1} \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}I_p \end{bmatrix}^T \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} \\ &= (\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

The solution above is the solution based on Exercise 3.12. It is also possible to derive the solution by straight forward differentiation. The above solution does, however, provide the possibility to compute the ridge regression estimate whenever you have access to ordinary least squares regression by augmentation of the dataset.

**Lasso**

If $J(\beta) = \sum_{i=1}^p |\beta_i| = ||\beta||_1$ the penalized estimation method is known as *lasso* = least absolute shrinkage and selection operator.

We need to optimize

$$(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\sum_{k=1}^p |\beta_k|.$$

The function $J$ is *convex* with $J(\beta) \to \infty$ for $||\beta|| \to \infty$. If there is a unique least squares solution there is a unique minimum $\hat{\beta}^{\text{lasso}}$.

This is a convex, but non-differentiable optimization problem.

**Restricted Estimation**

If $C \subseteq \mathbb{R}^p$ the *restricted estimator* is the estimator

$$\hat{\beta}^C = \operatorname*{argmin}_{\beta \in C} (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta).$$

The optimization problem is nicest if $C$ is convex. A well known situation is when $C$ is a subspace parameterized as

$$C = \{A\delta \mid \delta \in \mathbb{R}^q\}$$

where $A$ is a $p \times q$ rank $q$ matrix. The solution is

$$\hat{\beta}^C = (A^T\mathbf{X}^T\mathbf{X}A)^{-1}A^T\mathbf{X}^T\mathbf{y}.$$

**Duality**

If $J : \mathbf{R}^p \to [0, \infty)$ is a function we can define the *sub-level sets*

$$C_J(s) = \{\beta \mid J(\beta) \leq s\}.$$

If $J$ is convex then $C_J(s)$ is convex for all $s$. The function

$$\lambda \to s(\lambda) := J(\hat{\beta}^{\lambda J})$$

is typically a continuous, strictly decreasing function with $s(\lambda) \to 0$ for $\lambda \to \infty$ mapping $[0, \infty)$ onto $(0, s(0)]$.

$$\boxed{\hat{\beta}^{\lambda J} = \hat{\beta}^{C_J(s(\lambda))}}$$

This gives a dual viewpoint on the penalized estimator as a restricted estimator and vice versa for level set restrictions.

For the ridge regression the exercise Theo.2 deals with showing the above result based on the explicit representation of the ridge regression estimator. A similar result can be shown for lasso, but it is more difficult as we have no explicit representation. In general, the result belongs to the theory of constraint optimization. However, one has to be extremely careful here as we do not want to assume differentiability of $J$ since this rules out lasso. Anyway, this author is not sufficiently familiar with the theory of constraint optimization to formulate the correct mathematically rigorous version of the vague formulation above. Hence we keep a "typical" here and do some verification in a case by case manner. Unless the reader is quite familiar with the concepts of Lagrange multipliers etc. the introduction of such concepts does not seem to clarify much at this point.

**Figure 3.11 – Ridge and Lasso as Restricted Estimators**

Ridge regression (right) is a constraint optimization problem over a set with a smooth boundary. Lasso (left) is a constraint optimization problem over a set where the boundary has corners. The corners give lasso the selection ability.

**Duality**

Penalization can be viewed as an implicit *model restriction* – but in a *data dependent* way through $s(\lambda)$.

The parameterized family of solutions $(\hat{\beta}^{C_J(s)})_{s\in(0,s(0)]}$ is identical to the family $(\hat{\beta}^{\lambda J})_{\lambda\geq 0}$.

For lasso, optimization of

$$(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)$$

subject to $||\beta||_1 \leq s$ is a *quadratic* optimization problem subject to *linear* constraints, which is a classical numerical problem.

**Bridge Regression and The Elastic Net**

Generalizations include

- Bridge regression

$$J(\beta) = \sum_{i=1}^{p} |\beta|^q$$

  for $q \in (0,\infty)$
  - $q = 2$ is ridge regression
  - $q = 1$ is lasso
  - $q < 1$ is non-convex
  - $q \to 0$ is best subset selection
- The elastic net

$$J(\beta) = \alpha\beta^T\beta + (1-\alpha)\sum_{i=1}^{p} |\beta|$$

  for $\alpha \in [0,1]$.

**Figure 3.12 – Bridge and Elastic Net**

For $q \leq 1$ gives corners and has the *selection property*. For $q < 1$ we have a non-convex problem, $q \to 0$ results in best subset selection. With $q = 1$ we get selection as well as convexity. The elastic net has selection but more convexity.

**Figure 3.10 – Lasso Profiles and LARS**

The recent algorithm *lar* = least angle regression, or rather *lars* = lar with lasso modification, computes in one run *all* lasso estimates.

The path $\hat{\beta}^{\text{lasso},s}$ for $s$ varying is piecewise linear – here $s$ is scaled by $s(0)$ so $s \in (0,1]$.

Even though the relation between the restricted optimization problem and the penalized optimization problem is data dependent, their entire paths coincide and would both be piecewise linear. All that would change from the penalized version to the restricted version is a monotone transformation of the 1st axis on the path plot.

**Derived Input Methods**

A third idea is to derive a new set of predictors $\mathbf{z}_1, \ldots, \mathbf{z}_M \in \mathbb{R}^N$ for $M \leq p$ from $\mathbf{X}$, replace $\mathbf{X}$ by

$$\mathbf{Z} = [\mathbf{z}_1 \ldots \mathbf{z}_M]$$

and regress using these *derived input*.

- *Principal components regression (PCR).* With $\mathbf{X} = \mathbf{UDV}^T$ the singular value decomposition take $\mathbf{z}_i = \mathbf{u}_i$.
- *Partial least squares (PLS).* Popular in chemometrics. Includes $\mathbf{y}$ in the selection of the directions as opposed to PCR.

PCR chooses the directions *disregarding* $\mathbf{y}$. We close our eyes and hope these directions matter.

Neither PCR nor PLS are invariant to scaling. Either the measurements should be standardized or measured on a common scale.

In Theo.1 it is shown that the first principal component is the direction where we get the least uncertain estimate of the regression estimate measured in terms of variance. However, variance comparison across different parameters make little sense if the parameters are not on a common scale. The variance measured in meters is bound to be smaller than the variance measured in centimeters.

**Figure 3.16 – Comparisons**

The *forward stepwise* algorithm is greedy and achieves rapid improvement of the MSE for the parameter $\beta$.

LAR and lasso catches up later and ultimately outperforms forward stepwise.

**Figure 3.18 – Comparisons**

For a 2 dimensional parameter we can illustrate how the chosen parameters behave for different methods and different choices of selection/regularization.

Note that only ridge and lasso provide estimates on the entire curve plottet. The other three methods provide only one alternative to the least squares estimate (4,2).