Subset Selection: Which variables to keep and which to drop?

Why it's a difficult task? Can we just select variables based on their p-values in the `R` output, e.g., drop all variables which are not significant at 5%?

Subset Selection: Best Subset

1. score each model (model $=$ subset of variables)
2. design a search algorithm to find the optimal one.

Model selection criteria/scores for linear regression often take the following form

$$\text{Goodness-of-fit} + \text{Complexity-penalty}.$$

The 1st term is an increasing function of RSS, and the 2nd term an increasing function of p (the number of non-intercept variables).[a]

[a] Intercept is always included. You can count the intercept in p or not; It doesn't make any difference. From now on, $p =$ number of non-intercept variables.

Popular choices of scores:

- Mallow's C_p: $\text{RSS} + 2\hat{\sigma}^2_{\text{full}} \times p$ [a]
- AIC: $-2\text{loglik} + 2p$ [b]
- BIC: $-2\text{loglik} + (\log n)p$

Note that when n is large, adding an additional predictor costs a lot more in BIC than AIC. So AIC tends to pick a bigger model than BIC. C_p performs similar to AIC.

---

[a] $\hat{\sigma}^2$ is estimated from the full model (i.e., the model with all the predictors).

[b] In the context of linear regression with normal errors, we can replace -2loglik by $\log \text{RSS}$.

Mallow's C_p

- Recall the decomposition of the training and test error.

$$\begin{aligned}\mathbb{E}[\text{Train Err}] &= (\text{Unavoidable Err}) - p\sigma^2 + \text{Bias} \\ \mathbb{E}[\text{Test Err}] &= (\text{Unavoidable Err}) + p\sigma^2 + \text{Bias}\end{aligned}$$

- So Test Err $\approx$ RSS $+ 2p\sigma^2$, which is known as Mallow's C_p.