

Subset selection

→ find the "best" model among all possible models that we can fit using the available variables

$$y \quad X \quad (x_1, \dots, x_p) \quad 2^p$$

Goals

→ we want a model that can explain as much as possible the phenomenon of interest

→ the more the variables, the better

lower $RSS(p)$

higher R^2

→ we want a model that has as few variables as possible is as simple as possible

→ parsability

→ interpretability

→ generalizability

→ we have a trade-off between goodness-of-fit and generalizability → bias-variance trade-off (tomorrow)

→ we want to keep variables that has a ~~real~~ ^{systematic} effect (p-value $\leq \alpha$)

→ we want to remove those that has little effect or no-systematic effect (p-value $> \alpha$)

→ for today: based on the p-value of the related regression coefficient

→ in the previous two lectures we used a model-selection technique called "forward selection":

→ we start with the null model $y = \beta_0 + \varepsilon$

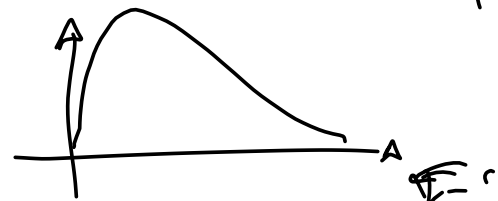
→ we add step-by-step new variables to the model, until no variable helps anymore improving the model

$$y = \beta_0 + \beta_1 \text{engine-size} + \varepsilon$$

$$y = \beta_0 + \beta_1 \text{engine-size} + \beta_2 \text{fuel-type} + \varepsilon$$

⋮

all p-values > α



Alternative: "backward elimination", that works the other way around

→ start with the full model

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$$

→ remove step-by-step those variables that do not help

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_{k-1} x_{k-1} + \beta_{k+1} x_{k+1} + \dots + \beta_p x_p + \varepsilon$$

until all variables have a significant effect

(all p-values < α)

we can also combine these two procedures:

- stepwise selection: forward selection in which it is allowed at each step to remove a variable previously added
- stepback selection: backward elimination in which at each step one is allowed adding a variable previously removed

These procedures can be seen as greedy approximations of the "best sub set" selection:

- fit all 2^p models
- find the best one → $\hat{R}^2$, AIC

In addition (more relevant for high-dimensional data regression than the current "classical" settings): forward-stagewise

- similarly to forward selection but, at each step, we do not update the regression coefficients of the variables already in the model