# Instuderingsfrågor – Kapitel 6.1 och 6.2

12 stycken frågor totalt.

## Kapitel 6.1 – Modellval

1. Logiken bakom mått såsom Cp, AIC, BIC och Adjusted R2 är:   
   *“We can indirectly estimate test error (test MSE) by making an adjustment to the training error to account for the bias due to overfitting.”*

* Förklara vad detta innebär.
* Behöver vi använda valideringsdata när vi använder något av måtten Cp, AIC, BIC eller Adjusted R2 ?

1. Mått såsom Cp, AIC, BIC och Adjusted R2 var smidigt att använda förr istället för ”validation set” metodiken. Varför?
2. AIC och BIC definieras enligt nedanstående två formler där man vill ha den modell som har lägst värde:   
   En bild som visar Teckensnitt, text, typografi, vit

   Automatiskt genererad beskrivning

En bild som visar Teckensnitt, typografi, text, kalligrafi

Automatiskt genererad beskrivning

* Notera, i formeln för BIC så står det log(n) men det är den naturliga logaritmen (ln) som menas.
* n är antalet observationer och d är antalet variabler i modellen.

Vilken av de två måtten, AIC och BIC, väljer generellt sett mindre modeller. Varför?

1. Förklara algoritmen nedan för ”Best subset selection”

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

1. Var är ett potentiellt problem med ”Best Subset Selection” när man har många variabler?
2. Vad är lika med?   
   - Ställer Antonio denna frågan för att han är nostalgisk efter kombinatorik som vi gjorde i statistik kursen eller finns det en annan anledning?
3. I steg 3 av algoritmen, varför kan vi inte välja den modell med lägst RSS eller ekvivalent högst R2 ?   
   Hint: När man tränar en regressionsmodell, vad är det som minimeras?
4. Förklara algoritmen nedan för ”Forward stepwise selection”:

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

1. Antag att ett dataset har p = 3 variabler (X1, X2, X3). Den bästa modellen som har en variabel innehåller X1 medan den bästa modellen med två variabler innehåller X2, X3, kommer ”Forward stepwise selection” modellen kunna rekommendera den bästa modellen som innehåller två variabler? Varför/Varför inte?
2. Förklara algoritmen nedan för ”Backward stepwise selection”:

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

## Kapitel 6.2 – Regularisering

1. I korthet, vad är intuitionen bakom ”Ridge regression” och ”LASSO regression”? Varför kan det vara bättre att ”begränsa flexibiliteten” för en regressions modell?
2. *”The lasso performs variable selection”*, i korthet vad menas med det påståendet?