

# Datorlaboration 1

Josef Wilzén

August 19, 2025

732G57 Maskininlärning för statistiker

## Allmänt

Datorlaborationerna kräver att ni har R och Rstudio installerat.

- **ISL**: An Introduction to Statistical Learning,
  - Boken: [länk](#)
  - R-kod till labbar: [länk](#)
  - Dataset: [länk](#) och [länk](#)

Notera att ni inte behöver göra alla delar på alla uppgifter. Det viktiga är att ni får en förståelse för de olika principerna och modellerna som avhandlats. Dessa uppgifter ska inte lämnas in, utan är till för er övning.

## R-kod

- R-paketet `glmnet` implementerar Ridge, LASSO och elastic-net: bra intro [här](#) och [här](#).
- Kodmanual för kursen 732G12 finns [här](#). Notera att denna inte är uppdaterad för detta års omgång.

## Frivillig repetition: Logistik regression

1. Gå igenom laborationerna 4.7.1 och 4.7.2 i **ISL**.

## 1 Modelval och Cross-Validation

1. Gå igenom laborationerna 5.3.1, 5.3.2, 5.3.3 i **ISL**.
2. **ISL** 5.4 Exercises Conceptual: Gör uppgift 3)

## 2 Variabelselektion för linjära modeller

1. Gå igenom laborationen 6.5.1 i **ISL**.
2. Gå igenom laborationen 6.5.2 i **ISL**.
3. Använd Hitters-data och gå igenom koden här för att testa adaptive lasso.
4. Använd Hitters-data och testa elastic net regression. Detta görs genom att ändra argumentet `alpha=` till ett värde mellan 0 och 1 i funktionerna `glmnet` och `cv.glmnet`. Testa minst två olika värden på `alpha`, och jämför med tidigare resultat.
5. **ISL 6.6 Exercises Conceptual**: Gör uppgift 2) och 4)

## 3 Variabelselektion för logistik regression

1. Gå till UC Irvine Machine Learning Repository, länk här. Klicka på “VIEW DATASETS”
2. Ni ska nu välja ut ett dataset här som passar för klassificering. Notera att ni kan filtrera på olika kategorier till vänster, välj “Classification”. Ni kan sortera och filtera dataset baserat på antal observationer och antal variabler.
3. Välj ett dataset som intresserar er, dock ej för stort, då det kan resultera i långsamma beräkningar. Nedan följer några förslag, men välj gärna andra:
  - (a) Divorce Predictors data set Data Set
  - (b) Wine Data Set
  - (c) Heart failure clinical records Data Set
4. Försök förstå (kort) vad datasetet handlar om och vad de olika variablerna innebär. Kolla om responsvariabeln är binär (logistic regression) eller nominell (multinomial regression), detta avgör vilken typ av modell ni ska använda.
5. Ladda ner datasetet och importera till R. Notera att många dataset kommer i något komprimerat format.
6. Genomför nödvändig datahantering, tex: kolla efter NA, skala om variabler, enkla plottar beskrivande statistik mm. Lägg inte för lång tid på detta. Om vissa variabler är svåra att förstå eller har många saknade värden, uteslut dem. Notera: `glmnet` behöver ha kategoriska variabler som lämpliga indikatorvariabler. Tips: använd funktionen `model.matrix()`.
7. Skapa ett valideringsset från ert dataset (vissa dataset har en förbestämt testset, användet det i så fall som valideringsset).

- (a) Se till att alla klasser för responsvariabeln finns representerade i valideringsset
  - (b) `createDataPartition()` i från paktet `caret` kan skapa träning-validering delning. Se kodmanualen.
8. Nu ska ni bygga en prediktiv modell (logistik regression) som kan prediktera observationer i valideringsdata. Se guiderna för `glmnet` för ridge/lasso tillsammans med logistik regression. Testa följande metoder för modellval:
- (a) Manuellt välja en delmängd av alla variabler och skatta som “vanligt” med `glm()` eller liknade funktion. Lägg inte för lång tid på att välja variabler.
  - (b) Forward selection
  - (c) Ridge regression med korsvalidering för  $\lambda$ . Tips: `cv.glmnet()`.
  - (d) LASSO regression med korsvalidering för  $\lambda$
  - (e) Elasticnet regression, med  $\alpha = 0.5$  och korsvalidering för  $\lambda$ .
9. Räkna ut klassificeringsfelet (misclassification rate) för valideringsdata. Vilken model var bäst? Undersök om det är lätt att identifiera vilka variabler som är viktiga för modellerna. För den bästa och näst bästa modellen:
- (a) Beräkna förväxlingsmatrisen för valideringsdata. Kod: här.
  - (b) Beräkna sensitivitet och specificitet för testdata.
  - (c) Vad skiljer den bästa och näst bästa modellen när det gäller klassificeringen? Är det stor eller liten skillnad?
  - (d) Vilka variabler var viktiga för bästa och näst bästa modellen när det gäller prediktion? Är det någon skillnad?

## 4 Mer övningsuppgifter

Nedan följer fler övningsuppgifter.

- Kapitel 5.4: Exercises Applied: 5), 7), 8)
- Kapitel 6.6: Exercises Applied: 8), 10)