

# Föreläsning 2 - Variabelselektion och regularisering

Josef Wilzen

2022-08-23

# Outline

- 1 Intro
- 2 Modelval
- 3 Generaliserade linjära modeller
- 4 Modelval för linjär regression

# Intro

- Kursplanen: [länk](#)
- Labbar

## Bias, varians, brus

$$y = f(x) + \varepsilon \quad E[\varepsilon] = 0 \quad V[\varepsilon] = \sigma^2$$

$$\hat{y} = \hat{f}(x_{test})$$

Förväntad test MSE:

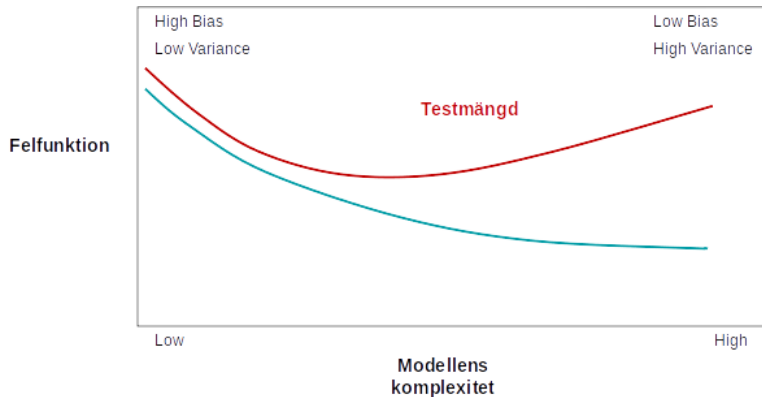
$$E \left[ y_{test} - \hat{f}(x_{test}) \right]^2 = V[\varepsilon] + V \left[ \hat{f}(x_{test}) \right] + Bias \left[ \hat{f}(x_{test}) \right]^2$$

- Brusvariens  $V[\varepsilon]$ : irreducibel brus
- Modellens varians  $V \left[ \hat{f}(x_{test}) \right]$ : Hur mycket kommer  $\hat{f}$  att ändras när vi byter dataset
- Modellens skewhet  $Bias \left[ \hat{f}(x_{test}) \right]$ : Systematisk skewhet eller modelleringsfel i modellen

# Bias, varians, brus

## Bias-variance-trade-off

- Vi vill ha
  - ▶ Lågt bias
  - ▶ Låg varians



# Modellval

- Vi vill ha bra **generalisering** hos modellen
  - ▶ Modellen ska fungera bra på nya liknande dataset!
- Komplexa modeller överanpassar lätt
- “Komplexitet” betyder olika saker för olika modeller
  - ▶ Linjära modeller: fler variabler, interaktioner, transformationer av variabler
  - ▶ Neurala nätverk: bred och djup
  - ▶ Trädmodeller: djup

# Regularisering

- Metoder för att undvika överanpassning
  - ▶ Hindra modellerna att bli för komplexa
  - ▶ Detta ger förhoppningsvis bättre generaliseringsfel
  - ▶ **Mycket viktigt tema inom maskininlärning**
  - ▶ Betyder olika saker för olika metoder
- Recept:
  - ▶ Komplex modell + regularisering = en bra modell

# Regression och klassificering

- Regression:  $y$  är kontinuerlig, bruset  $\varepsilon$ :
  - ▶ Antas ofta vara normalfördelat
  - ▶ Alt:  $t$ , Gamma, Log-normal
  - ▶ Frekvenser: Poisson, Negativ binomial
- klassificering:  $y$  är kategorisk med 2 eller fler utfall
  - ▶ Binär: logistik/probit regression
  - ▶ Fler klasser: multinomial logistik/probit regression
  - ▶ Fler metoder senare i kursen



# Förväxlingsmatris

		Predikterad klass	
		Class = 1	Class = 0
Sann klass	Class = 1	$f_{11}$	$f_{10}$
	Class = 0	$f_{01}$	$f_{00}$

- Precision:

$$P = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

- Felkvot (error rate):

$$E = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

# Specificitet och sensitivitet

		Predikterad klass	
		Class = 1	Class = 0
Sann klass	Class = 1	$f_{11}$	$f_{10}$
	Class = 0	$f_{01}$	$f_{00}$

- Sensitivitet (=true positive rate, recall):

$$= \frac{f_{11}}{f_{11} + f_{01}}$$

- Specificitet (=true negative rate):

$$= \frac{f_{00}}{f_{10} + f_{00}}$$

- Dessa mått är klassspecifika. Dessa formler betecknar klass 1.

# Specificitet och sensitivitet

		Predikterad klass	
		Class = 1	Class = 0
Sann klass	Class = 1	$f_{11}$	$f_{10}$
	Class = 0	$f_{01}$	$f_{00}$

- Sensitivitet (=true positive rate, recall):

$$= \frac{f_{11}}{f_{11} + f_{01}}$$

- Specificitet (=true negative rate):

$$= \frac{f_{00}}{f_{10} + f_{00}}$$

- Dessa mått är klassspecifika. Dessa formler betecknar klass 1.

# F-score

- F-score är det harmoniska medelvärdet av precisionen och sensitivitet (recall)
- $F_{\beta} = (1 + \beta^2) \frac{\text{precision} \cdot \text{sensitivitet}}{(\beta^2 \cdot \text{precision}) + \text{sensitivitet}}$
- Vanligt med  $\beta = 1$ :  $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{sensitivitet}}{\text{precision} + \text{sensitivitet}}$
- Ligger mellan 0 och 1, högre värden är bättre
- Andra värden: 0.5 och 2
- Beräknas klassvis

# Generaliserade linjära modeller

Generaliserade linjära modeller (GLM):

- $y_1, y_2, \dots, y_n$ : oberoende från sannolikhetsfördelning från exponentialfamiljen
- Linjär prediktor:  $X\beta$
- Länkfunktion som kopplar den linjära prediktorn till medelvärdet  $\mu$

$$g(\mu) = X\beta$$

# Generaliserade linjära modeller

- Generaliserar linjär regression till andra reponsvariabler
  - ▶ Andra likelihoodfunktioner
- Reponsvariabler
  - ▶ Kontinuerlig: normal, t, gamma, log-normal
  - ▶ Binär: Logistisk regression, probit regression
  - ▶ Nomiell: Multinomiell logistisk regression
  - ▶ Frekvensdata: Poission regression, Negativ binomialfördelning

# Linjär regression

- Likelihood: Normal
- Länkfunktion: identitetsfunktionen
- Skattas med genom att minimera:

$$RSS = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{i,j} \right)^2$$

# Logistisk regression

- Anta binär  $y$ .  $p = P(y = 1)$
- Odds:  $0 \leq \frac{p}{1-p} < \infty$ , Log odds = logit:  $-\infty < \log\left(\frac{p}{1-p}\right) < \infty$
- Logistisk regression antar att log oddset för  $P(y = 1)$  beror linjärt på de förklarande variablerna

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \Leftrightarrow$$

$$\begin{aligned} p &= \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p))} \\ &= \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)} \end{aligned}$$



# Logistisk regression

- Logistiska funktionen:

$$p = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p))}$$

- Prediktion: Om  $P(y = 1|X_{test}, \beta) > 0.5 \rightarrow$  klass 1
- Skattas med maximum likelihood (MLE), hitta  $\beta$  som maximerar:

$$l(\beta) = \prod_{i:y_i=1} P(y_i = 1|X_i, \beta) \prod_{i':y_{i'}=0} (1 - P(y_{i'} = 1|X_{i'}, \beta))$$

# Multinomiell regression

- Anta att  $y$  kan anta  $K$  olika värden
- Multinomiell logistisk regression använder länkfunktionen:

$$P(y_i = k) = \frac{\exp(\beta_{0,k} + \beta_k^T x)}{\sum_{l=1}^K \exp(\beta_{0,l} + \beta_l^T x)}$$

- Kallas för **softmaxfunktionen**
- Notera att vi har:  $\beta_{0,l} + \beta_l^T x$  för varje klass  $l$ .  $\beta$  är en matris med storlek  $p \times K$
- Skattas med MLE.

# Modelval för linjär regression

Utgå från:  $y$  kontinuerlig med normal likelihood

Vi har ett antal förklarande variabler  $X = (x_1, \dots, x_p)$ . Vill vi hitta den delmängd som ger minst generaliseringsfel på testdata. Två alternativ:

- Välj ut en delmängd av variablerna och skatta med OLS
  - ▶ Best subset, Forward selection, Backward selection
- Behåll alla variablerna med sätt begränsningar på variablernas parameterrum (support)  $\rightarrow$  det ger mindre flexibilitet  $\rightarrow$  minskar risken för överanpassning
  - ▶ Ridge, lasso

# Utväderingsmått

- Indirekt skatta testfelet
  - ▶ Utgår från träningsmängden
  - ▶ Försöker minska den bias som uppstår när vi bara använder oss av träningsmängden och inte “all data”
- Direkt skatta testfelet
  - ▶ valideringsdata
  - ▶ korsvalidering

## Indirekt skatta testfelet

$$C_p = \frac{1}{n} (RSS + 2d\hat{\sigma}^2)$$

$\hat{\sigma}^2$ : skattas ofta från den fulla modellen

Straffar med:  $2d\hat{\sigma}^2$

Litet  $C_p \rightarrow$  litet testfel

$$\text{adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

Stort  $\text{adjusted } R^2 \rightarrow$  litet testfel

## Indirekt skatta testfelet

AIC, BIC och HQIC: baserad på MLE skattning av modeller. Låt  $\log(\hat{L})$  vara värdet på log-likelihoodfunktionen för optimala parametervärden.

$$AIC = 2k - 2\log(\hat{L})$$

$$BIC = k \cdot \log(n) - 2\log(\hat{L})$$

$$HQIC = k \cdot \log(\log(n)) - 2\log(\hat{L})$$

Låga värden  $\rightarrow$  litet testfel

## Indirekt skatta testfelet

För linjär regression

$$AIC = \frac{1}{n\hat{\sigma}^2} (RSS + 2d\hat{\sigma}^2)$$

$$BIC = \frac{1}{n\hat{\sigma}^2} (RSS + \log(n) d\hat{\sigma}^2)$$

$$HQIC = \frac{1}{n\hat{\sigma}^2} (RSS + \log(\log(n)) d\hat{\sigma}^2)$$

Linjär regression:  $C_p \propto AIC$

Hur hög kostnad för fler parametrar?

- $n = 8$

$$2 < \log(n)$$

- $n = 2000$

$$2 < \log(\log(n)) < \log(n)$$

- BIC och HQIC väljer mindre antal parametrar generellt. För stora datamängder så tenderar AIC att välja för stora modeller.

# Best subset

Notera det finns  $2^p$  modeller att undersöka! Ex:  $2^{20} = 1048576$

---

## Algorithm 6.1 *Best subset selection*

---

1. Let  $\mathcal{M}_0$  denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
  2. For  $k = 1, 2, \dots, p$ :
    - (a) Fit all  $\binom{p}{k}$  models that contain exactly  $k$  predictors.
    - (b) Pick the best among these  $\binom{p}{k}$  models, and call it  $\mathcal{M}_k$ . Here *best* is defined as having the smallest RSS, or equivalently largest  $R^2$ .
  3. Select a single best model from among  $\mathcal{M}_0, \dots, \mathcal{M}_p$  using cross-validated prediction error,  $C_p$  (AIC), BIC, or adjusted  $R^2$ .
- 

Från "An Introduction to Statistical Learning with Applications in R" av Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani



# Forward selection

---

## Algorithm 6.2 *Forward stepwise selection*

---

1. Let  $\mathcal{M}_0$  denote the *null* model, which contains no predictors.
  2. For  $k = 0, \dots, p - 1$ :
    - (a) Consider all  $p - k$  models that augment the predictors in  $\mathcal{M}_k$  with one additional predictor.
    - (b) Choose the *best* among these  $p - k$  models, and call it  $\mathcal{M}_{k+1}$ . Here *best* is defined as having smallest RSS or highest  $R^2$ .
  3. Select a single best model from among  $\mathcal{M}_0, \dots, \mathcal{M}_p$  using cross-validated prediction error,  $C_p$  (AIC), BIC, or adjusted  $R^2$ .
- 

Från "An Introduction to Statistical Learning with Applications in R" av Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Backward selection

---

**Algorithm 6.3** *Backward stepwise selection*

---

1. Let  $\mathcal{M}_p$  denote the *full* model, which contains all  $p$  predictors.
  2. For  $k = p, p - 1, \dots, 1$ :
    - (a) Consider all  $k$  models that contain all but one of the predictors in  $\mathcal{M}_k$ , for a total of  $k - 1$  predictors.
    - (b) Choose the *best* among these  $k$  models, and call it  $\mathcal{M}_{k-1}$ . Here *best* is defined as having smallest RSS or highest  $R^2$ .
  3. Select a single best model from among  $\mathcal{M}_0, \dots, \mathcal{M}_p$  using cross-validated prediction error,  $C_p$  (AIC), BIC, or adjusted  $R^2$ .
- 

Från "An Introduction to Statistical Learning with Applications in R" av Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Shrinkage

## Shrinkage (krympning)

- Hindra parameterarna från att bli “för stora”
  - ▶ Sätta begränsningar på parameterrummet
  - ▶ Ändra deras support (värdemängd)

## Vanligaste metoderna:

- Ridge:  $l^2$ -norm (Euclidean norm)
- LASSO:  $l^1$ -norm
- Standardisera era förklarande variabler först!

# Ridge regression

- Vanlig regression minimerar konstandsfunktionen

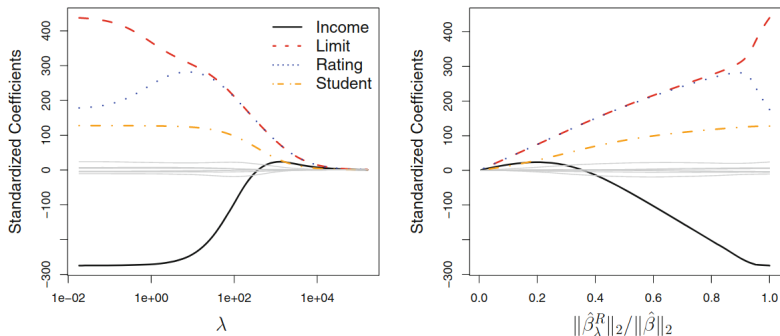
$$f(\beta) = RSS = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{i,j} \right)^2$$

- Ridge = OLS +  $l^2$ -norm på  $\beta$ :

$$\arg \min_{\beta} f(\beta) = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad \lambda \geq 0$$

- $\lambda$  är en **hyperparameter**.
- $\lambda \sum_{j=1}^p \beta_j^2$  = shrinkage penalty. Absolut stora värden på  $\beta$  "kostar mer". Hur mycket mer beror på  $\lambda$  (tänk prissättningen)
- Vad händer när  $\lambda \rightarrow 0$ ? eller  $\lambda \rightarrow \infty$ ?
- Notera att  $\beta_0$  inte påverkas av konstandsfunktionen

# Ridge regression



**FIGURE 6.4.** The standardized ridge regression coefficients are displayed for the **Credit** data set, as a function of  $\lambda$  and  $\|\hat{\beta}_\lambda^R\|_2 / \|\hat{\beta}\|_2$ .

Från "An Introduction to Statistical Learning with Applications in R" av Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Ridge regression

- $\lambda \rightarrow 0, \hat{\beta}_{ridge} \rightarrow \hat{\beta}_{OLS}$
- $\lambda \rightarrow \infty, \hat{\beta}_{ridge} \rightarrow 0$
- $\sum_{j=1}^p \beta_j^2 = \|\beta\|_2^2$ : kvadrerad  $l^2$ -norm
- $\sum_{j=1}^p \beta_j^2$ : definierar en hypersfär i det p-dimensionella rummet.
  - ▶  $p = 2$ : cirkel,  $p = 3$ : sfär
- Krymper: “alla lika mycket”
  - ▶ två korrelerade variabler: tar med båda, men krymper effektstorleken

# Lasso regression

- Vanlig regression minimerar konstantsfunktionen

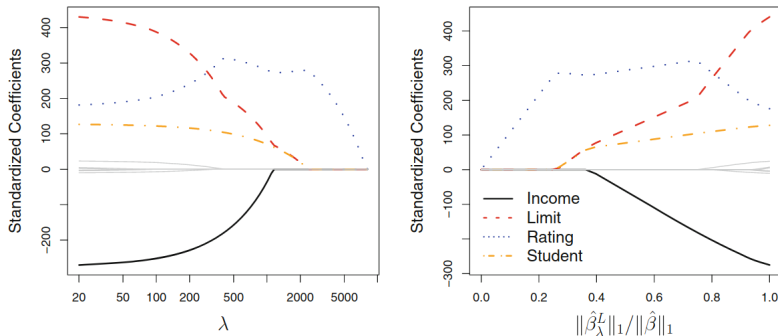
$$f(\beta) = RSS = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{i,j} \right)^2$$

- lasso = OLS +  $l^1$ -norm på  $\beta$ :

$$\arg \min_{\beta} f(\beta) = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad \lambda \geq 0$$

- $\lambda$  är en **hyperparameter**.
- $\lambda \sum_{j=1}^p \beta_j^2$  = shrinkage penalty. Absolut stora värden på  $\beta$  "kostar mer".
- lasso = least absolute shrinkage and selection operator

# Lasso regression



**FIGURE 6.6.** The standardized lasso coefficients on the **Credit** data set are shown as a function of  $\lambda$  and  $\|\hat{\beta}_\lambda^L\|_1 / \|\hat{\beta}\|_1$ .

Från “An Introduction to Statistical Learning with Applications in R” av Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani



# Lasso regression

- $\lambda \rightarrow 0, \hat{\beta}_{lasso} \rightarrow \hat{\beta}_{OLS}$
- $\lambda \rightarrow \infty, \hat{\beta}_{lasso} \rightarrow 0$
- $\sum_{j=1}^p |\beta_j| = \|\beta\|_1$ :  $l^1$ -norm
- $\sum_{j=1}^p |\beta_j|$ : definerar en polytop i det  $p$ -dimensionella rummet.
  - ▶  $p = 2$ : diamant,  $p = 3$ : polyeder
- Lasso: kan tvinga vissa  $\beta$  av bli 0  $\rightarrow$  variabelselektion, glesa (sparse) lösningar
  - ▶ ex: Låt  $p = 100$ , Då kan lasso göra att endast 20  $\beta$ -koefficienter är  $\neq 0$
- Två korrelerade variabler:
  - ▶ "tar en och kastar bort den andra"  $\rightarrow$  kan ge hög varians för testdata

# Adaptive lasso

- Låt varje  $\beta_j$  ha sin egen vikt  $w_j$  i regulariseringen

$$\arg \min_{\beta} f(\beta) = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^p w_j |\beta_j| \quad \lambda \geq 0$$

- Välj vikter

$$w_j = \frac{1}{\left| \hat{\beta}_{j,start} \right|^{\gamma}}$$

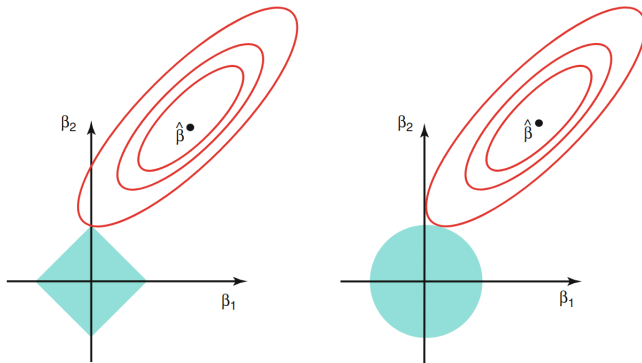
- Ofta så väljs  $\gamma = 1$ .
- Välj  $\hat{\beta}_{j,start}$  som  $\hat{\beta}_{j,OLS}$  eller  $\hat{\beta}_{j,ridge}$
- Stora värden på  $\hat{\beta}_{j,start}$  ger mindre vikter  $\rightarrow$  ger mindre krympning av  $\beta_j$
- Har visa teoretiska fördelar jämfört med vanlig lasso. Kräver dock att vi skattar  $\hat{\beta}_{j,start}$  i ett separat steg först.

# Ridge och lasso

Kan formuleras om:

- Minimera RSS med villkoret
  - ▶ Ridge:  $\sum_{j=1}^p \beta_j^2 \leq s$
  - ▶ Lasso:  $\sum_{j=1}^p |\beta_j| \leq s$
  - ▶ Varje  $\lambda$  motsvarar ett  $s$

# Ridge och lasso



**FIGURE 6.7.** Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions,  $|\beta_1| + |\beta_2| \leq s$  and  $\beta_1^2 + \beta_2^2 \leq s$ , while the red ellipses are the contours of the RSS.

Från "An Introduction to Statistical Learning with Applications in R" av Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Rigde vs lasso

- Båda skattas enkelt med `glmnet()` eller `cv.glmnet()`.
- Vilken som är “bäst” beror på kontexten
- Ridge passar för när  $y$  beror på de flesta förklarande variablerna
  - ▶ Många variabler och ungefär samma effektstorlek
- Lasso passar för när  $y$  beror på de ett fåtal förklarande variabler
  - ▶ Fåtal variabler med hög effektstorlek, resten nära 0
- Lasso kan ofta vara lättare att tolka
- Lättare att göra inferens för parametrarna för ridge
- Ofta får frågan avgöras empiriskt för specifika dataset

# Elasticnet regression

- Elasticnet kombinerar ridge och lasso
- Två hyperparametrar:  $\lambda$  och  $\alpha$

$$f(\beta) = RSS + \lambda \left[ (1 - \alpha) \sum_{j=1}^p \beta_j^2 + \alpha \sum_{j=1}^p |\beta_j| \right] \quad \lambda \geq 0, 0 \leq \alpha \leq 1$$

- $\lambda$ : generella styrkan på kostnaden
- $\alpha$ : relativa vikten på lasso och ridge,  $\alpha = 1$  ger lasso.

# Elasticnet regression

- Kan tvinga vissa  $\beta$  att bli 0  $\rightarrow$  variabelselektion, dock inte lika mycket som lasso
- Klarar av korrelerade/grupper av variabler bättre än lasso
- Nackdel: två hyperparameterar att specificera

# Regularisering

- Överanpassning är stort problem inom maskininlärning
- Regularisering:
  - ▶ Förekommer mycket ofta
  - ▶ Kan se olika ut beroende på modellklassen
- Ridge och Lasso
  - ▶  $\sum_{j=1}^p \beta_j^2 = \|\beta\|_2^2$ : kvadrerad  $l^2$ -norm
  - ▶  $\sum_{j=1}^p |\beta_j| = \|\beta\|_1$ :  $l^1$ -norm
  - ▶ Förekommer i många olika varianter



# Avslut

- Frågor? Kommentarer?
- Kurshemsidan
- Labben