Regularisering i regression

Mattias Villani

Statistiska institutionen
Stockholms universitet

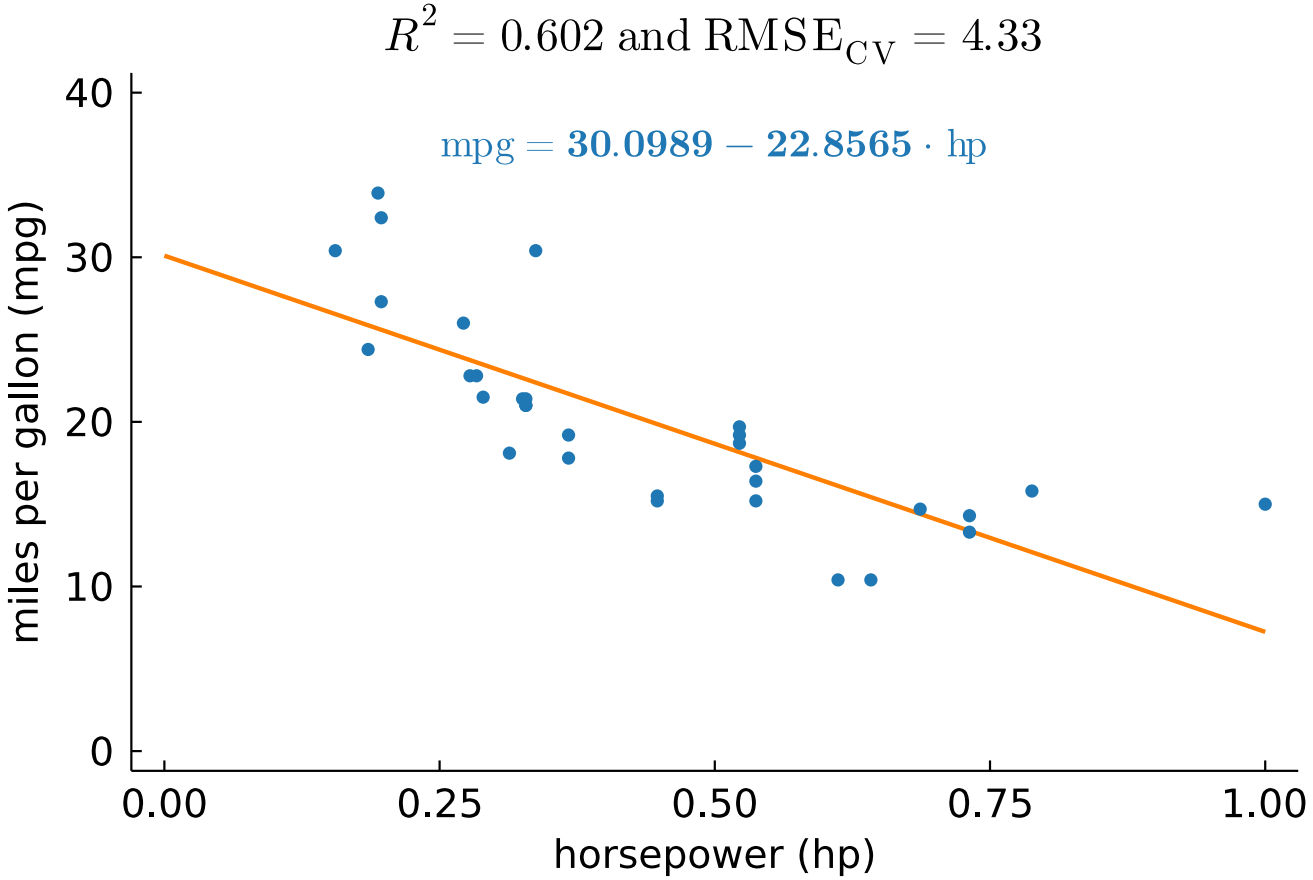
Institutionen för datavetenskap
Linköpings universitet


mattiasvillani.com


@matvil


mattiasvillani

mtcars data - linjär regression mot hp



Prognosförmåga på nya data - Korsvalidering

Fold 1

biltyp	mpg	hp
Hornet Sportabout	18.7	0.52
Fiat X1-9	27.3	0.20
Merc 450SL	17.3	0.54
Merc 450SLC	15.2	0.54
Merc 240D	24.4	0.19
Duster 360	14.3	0.73
Datsun 710	22.8	0.28
Ferrari Dino	19.7	0.52
Ford Pantera L	15.8	0.79
Pontiac Firebird	19.2	0.52
Toyota Corona	21.5	0.29
AMC Javelin	15.2	0.45
Camaro Z28	13.3	0.73
Fiat 128	32.4	0.20
Merc 280C	17.8	0.37
Lotus Europa	30.4	0.34
Cadillac Fleetwood	10.4	0.61
Chrysler Imperial	14.7	0.69
Mazda RX4	21	0.33
Volvo 142E	21.4	0.33
Mazda RX4 Wag	21	0.33
Merc 230	22.8	0.28
Toyota Corolla	33.9	0.19
Merc 280	19.2	0.37
Dodge Challenger	15.5	0.45
Lincoln Continental	10.4	0.64
Valiant	18.1	0.31
Honda Civic	30.4	0.16
Hornet 4 Drive	21.4	0.33
Merc 450SE	16.4	0.54
Maserati Bora	15	1.00
Porsche 914-2	26	0.27

Fold 2

biltyp	mpg	hp
Hornet Sportabout	18.7	0.52
Fiat X1-9	27.3	0.20
Merc 450SL	17.3	0.54
Merc 450SLC	15.2	0.54
Merc 240D	24.4	0.19
Duster 360	14.3	0.73
Datsun 710	22.8	0.28
Ferrari Dino	19.7	0.52
Ford Pantera L	15.8	0.79
Pontiac Firebird	19.2	0.52
Toyota Corona	21.5	0.29
AMC Javelin	15.2	0.45
Camaro Z28	13.3	0.73
Fiat 128	32.4	0.20
Merc 280C	17.8	0.37
Lotus Europa	30.4	0.34
Cadillac Fleetwood	10.4	0.61
Chrysler Imperial	14.7	0.69
Mazda RX4	21	0.33
Volvo 142E	21.4	0.33
Mazda RX4 Wag	21	0.33
Merc 230	22.8	0.28
Toyota Corolla	33.9	0.19
Merc 280	19.2	0.37
Dodge Challenger	15.5	0.45
Lincoln Continental	10.4	0.64
Valiant	18.1	0.31
Honda Civic	30.4	0.16
Hornet 4 Drive	21.4	0.33
Merc 450SE	16.4	0.54
Maserati Bora	15	1.00
Porsche 914-2	26	0.27

Fold 3

biltyp	mpg	hp
Hornet Sportabout	18.7	0.52
Fiat X1-9	27.3	0.20
Merc 450SL	17.3	0.54
Merc 450SLC	15.2	0.54
Merc 240D	24.4	0.19
Duster 360	14.3	0.73
Datsun 710	22.8	0.28
Ferrari Dino	19.7	0.52
Ford Pantera L	15.8	0.79
Pontiac Firebird	19.2	0.52
Toyota Corona	21.5	0.29
AMC Javelin	15.2	0.45
Camaro Z28	13.3	0.73
Fiat 128	32.4	0.20
Merc 280C	17.8	0.37
Lotus Europa	30.4	0.34
Cadillac Fleetwood	10.4	0.61
Chrysler Imperial	14.7	0.69
Mazda RX4	21	0.33
Volvo 142E	21.4	0.33
Mazda RX4 Wag	21	0.33
Merc 230	22.8	0.28
Toyota Corolla	33.9	0.19
Merc 280	19.2	0.37
Dodge Challenger	15.5	0.45
Lincoln Continental	10.4	0.64
Valiant	18.1	0.31
Honda Civic	30.4	0.16
Hornet 4 Drive	21.4	0.33
Merc 450SE	16.4	0.54
Maserati Bora	15	1.00
Porsche 914-2	26	0.27

Fold 4

biltyp	mpg	hp
Hornet Sportabout	18.7	0.52
Fiat X1-9	27.3	0.20
Merc 450SL	17.3	0.54
Merc 450SLC	15.2	0.54
Merc 240D	24.4	0.19
Duster 360	14.3	0.73
Datsun 710	22.8	0.28
Ferrari Dino	19.7	0.52
Ford Pantera L	15.8	0.79
Pontiac Firebird	19.2	0.52
Toyota Corona	21.5	0.29
AMC Javelin	15.2	0.45
Camaro Z28	13.3	0.73
Fiat 128	32.4	0.20
Merc 280C	17.8	0.37
Lotus Europa	30.4	0.34
Cadillac Fleetwood	10.4	0.61
Chrysler Imperial	14.7	0.69
Mazda RX4	21	0.33
Volvo 142E	21.4	0.33
Mazda RX4 Wag	21	0.33
Merc 230	22.8	0.28
Toyota Corolla	33.9	0.19
Merc 280	19.2	0.37
Dodge Challenger	15.5	0.45
Lincoln Continental	10.4	0.64
Valiant	18.1	0.31
Honda Civic	30.4	0.16
Hornet 4 Drive	21.4	0.33
Merc 450SE	16.4	0.54
Maserati Bora	15	1.00
Porsche 914-2	26	0.27

Träning
Test

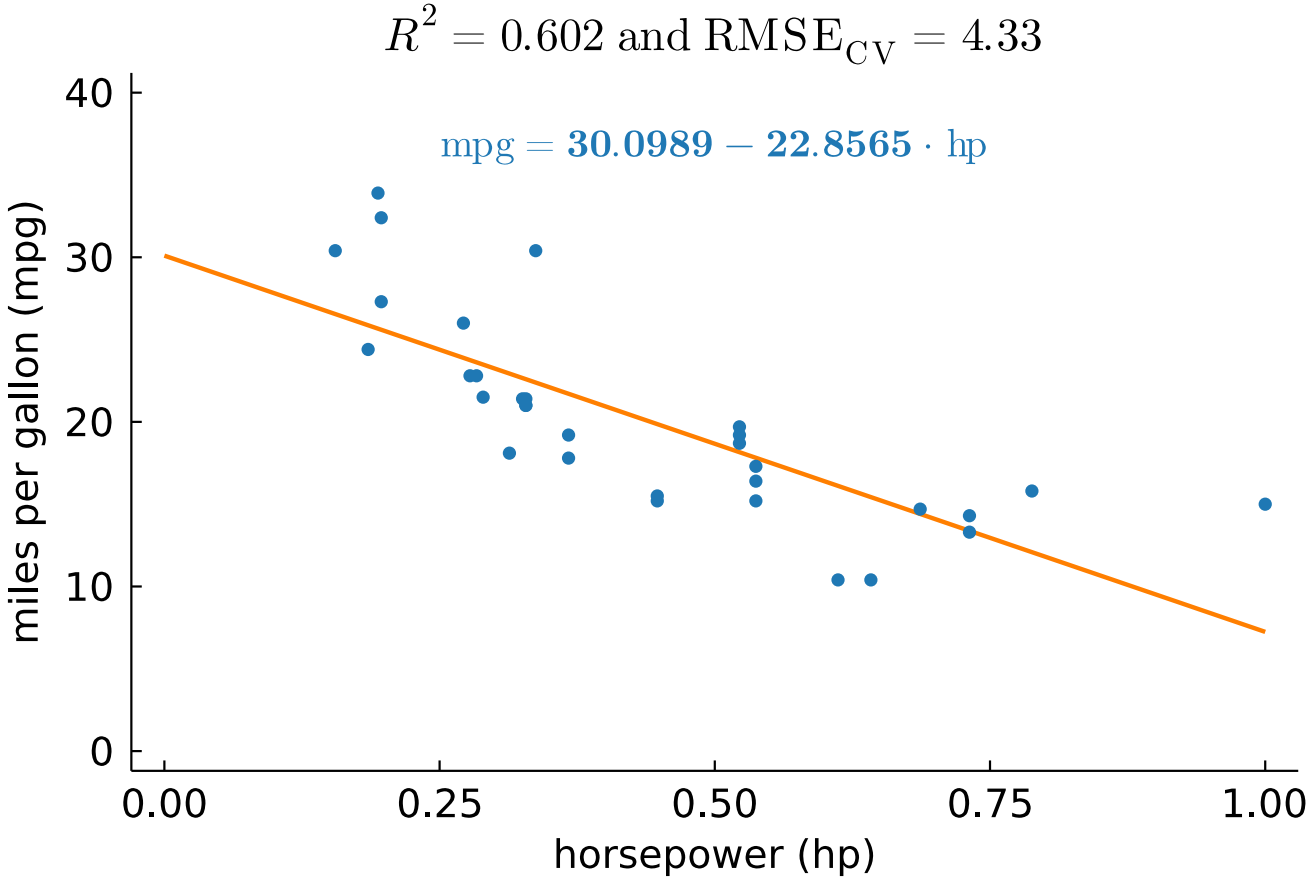
- **Prognosfel i testdata**

$$\text{MSE}_{CV} = \frac{\sum_{j=1}^{n_{\text{test}}} (y_j - \hat{y}_j^{\star})^2}{n}$$

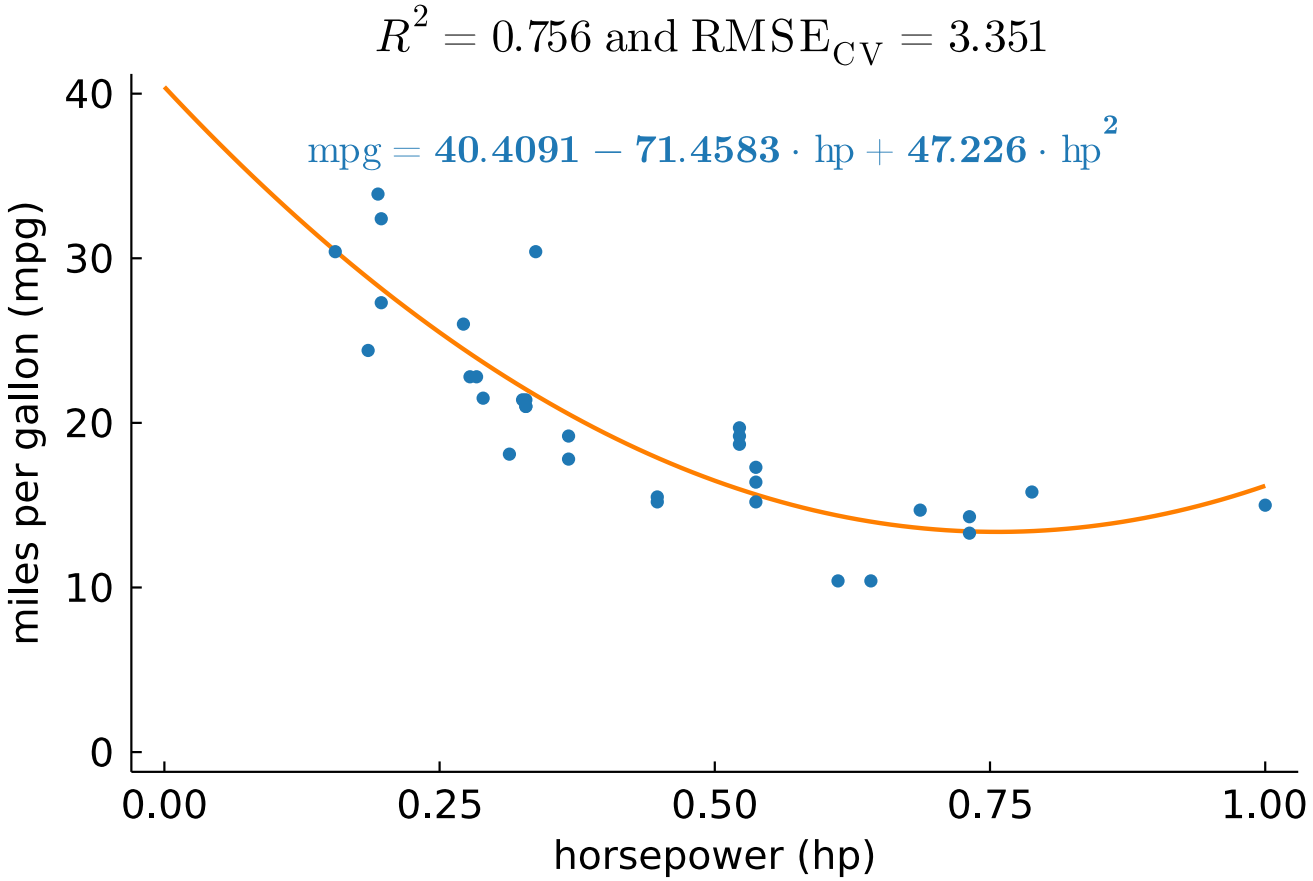
- Lättare att tolka **Root MSE**

$$\text{RMSE}_{CV} = \sqrt{\text{MSE}_{CV}}$$

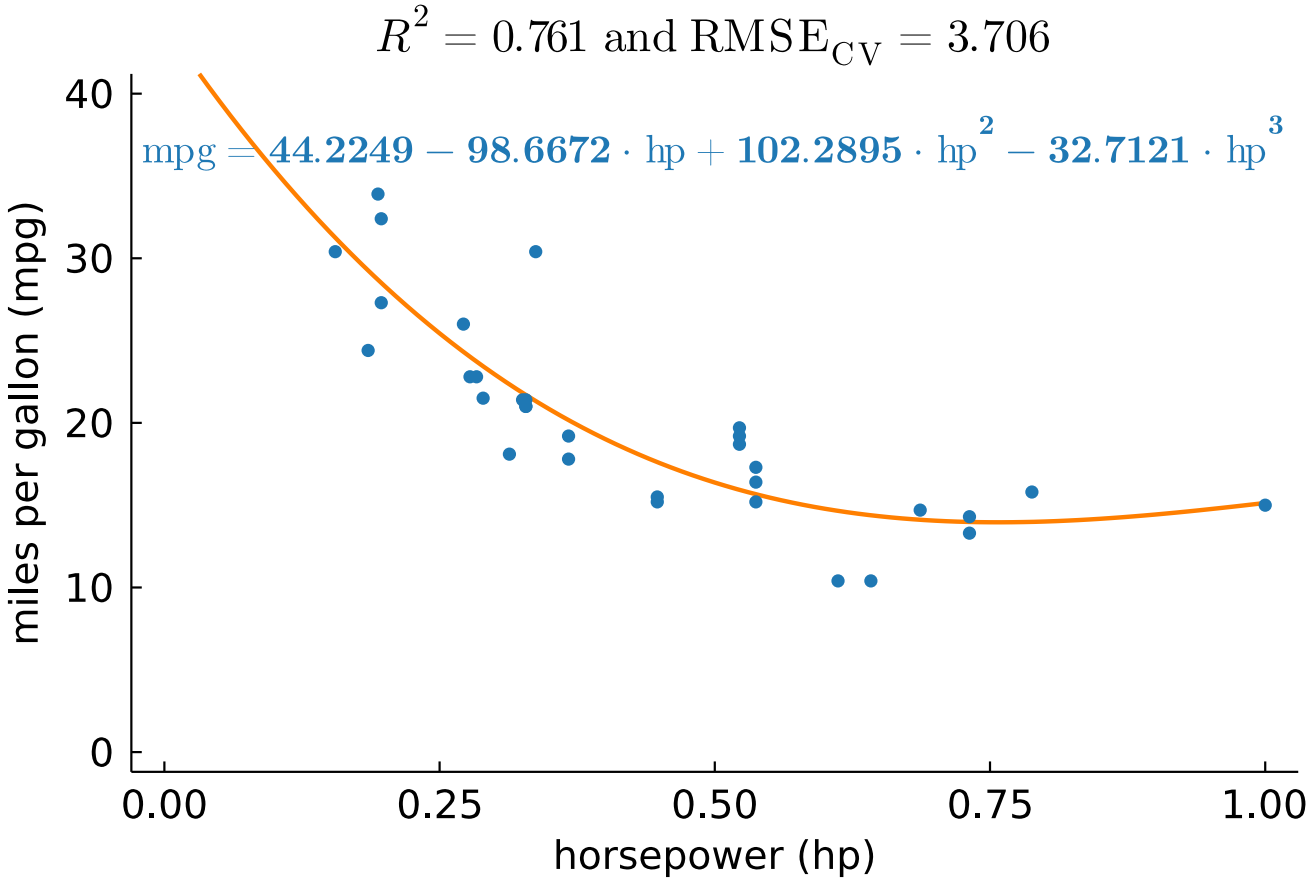
mtcars data - linjär regression mot hp



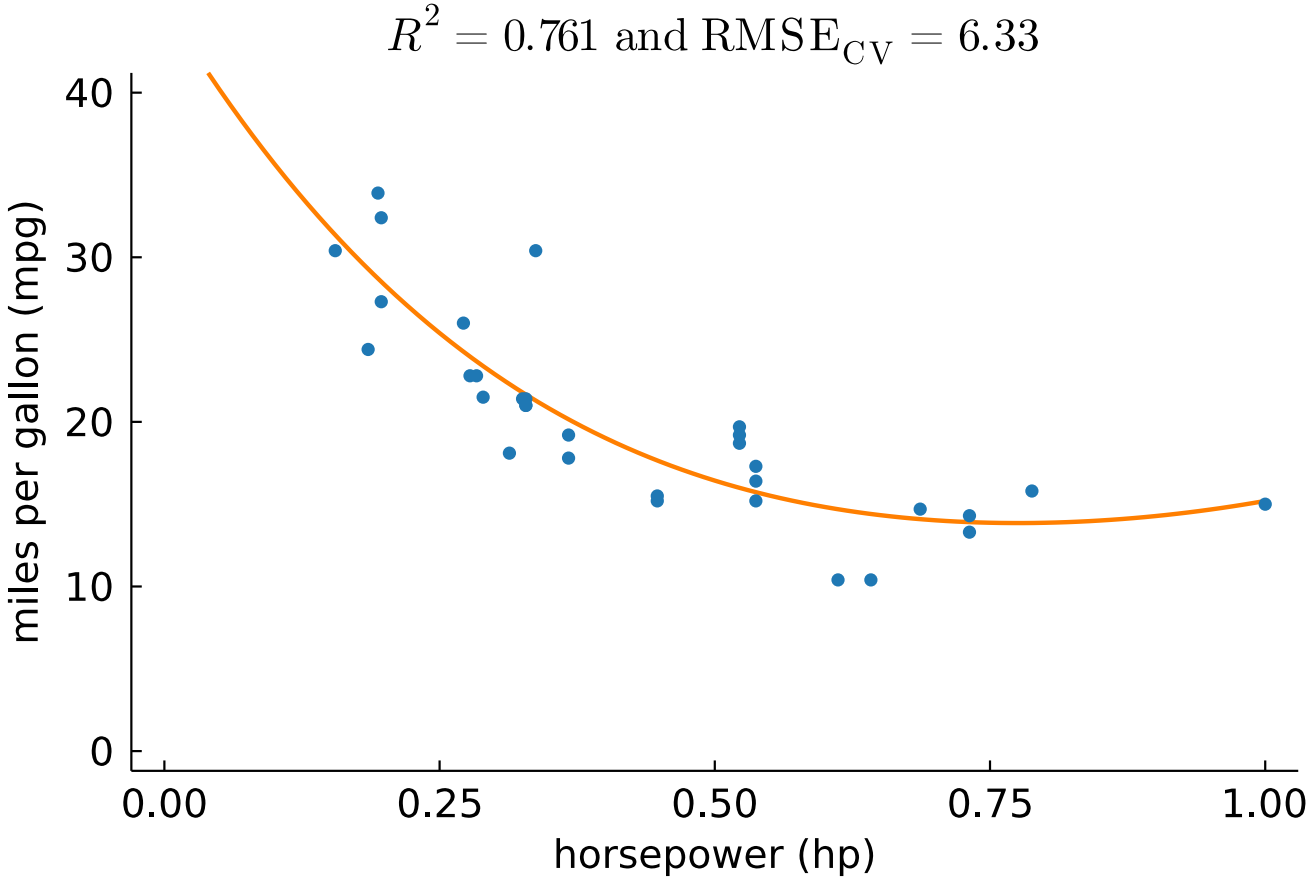
mtcars data - kvadratisk regression mot hp



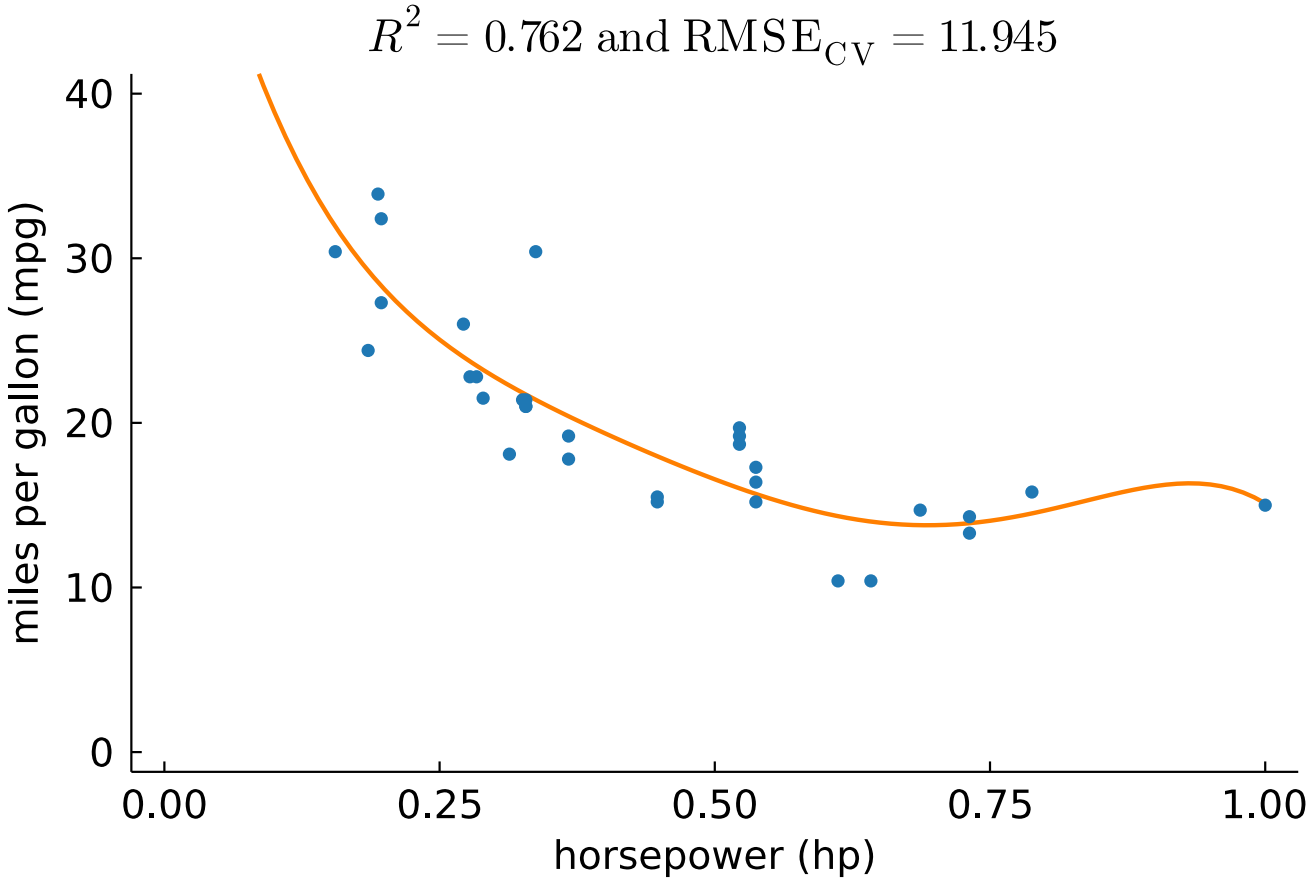
mtcars data - kubisk regression mot hp



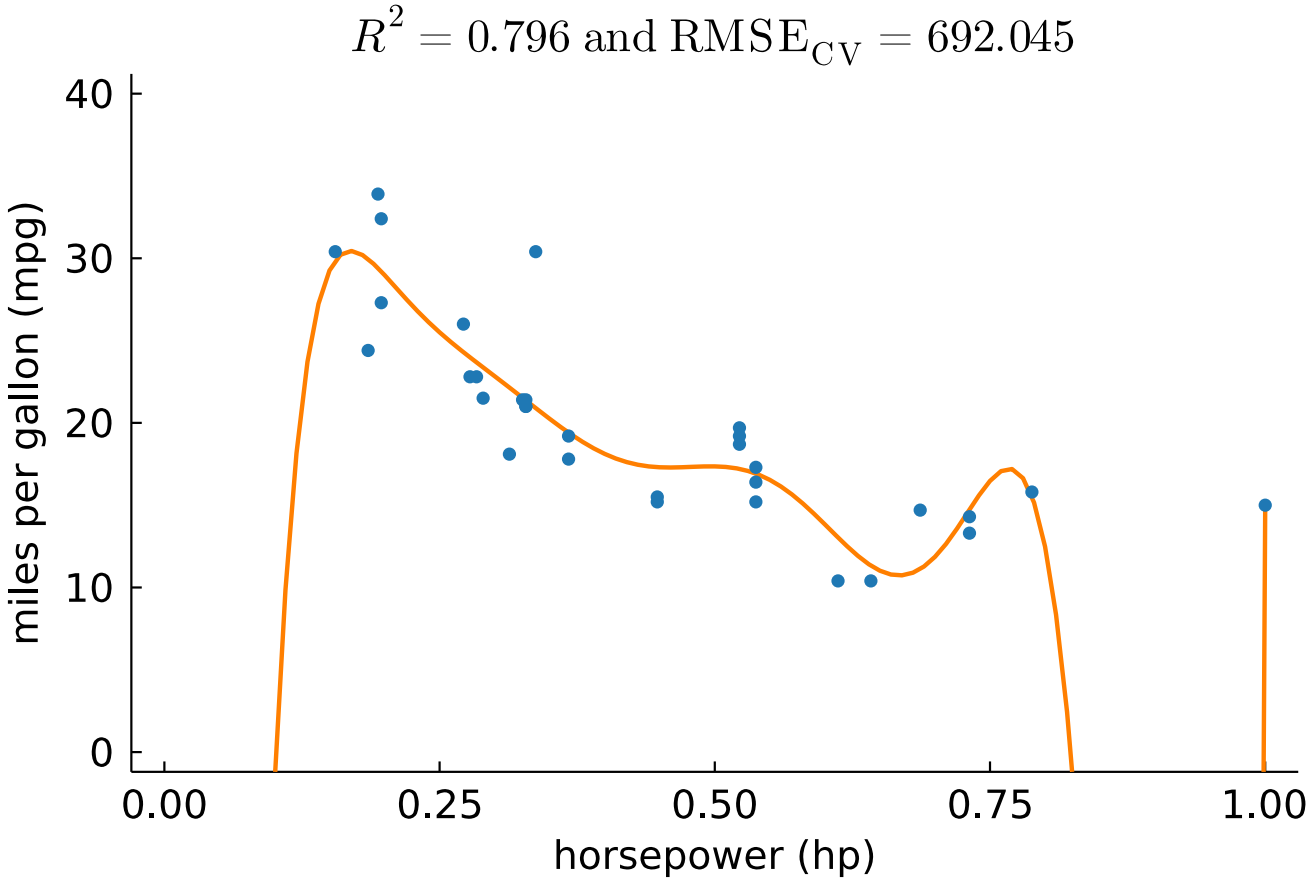
mtcars data - polynomregression ordning 4



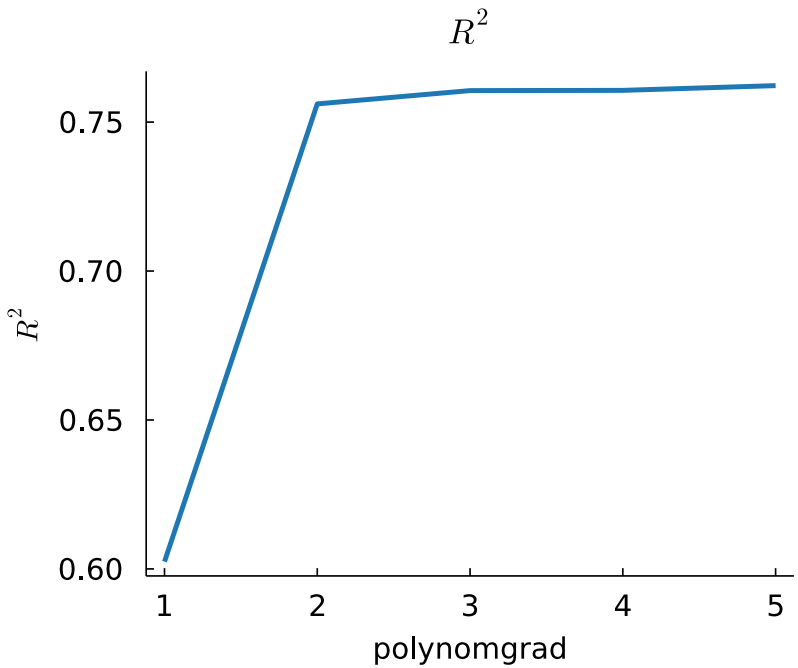
mtcars data - polynomregression ordning 5

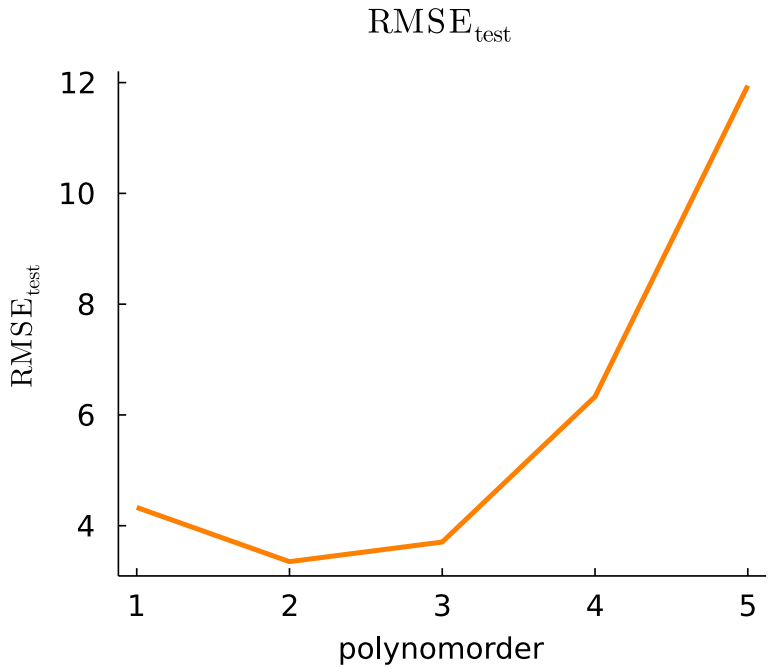


mtcars data - polynomregression ordning 10



mtcars data - R^2 och RMSE-CV($K = 4$)





L2-regularisering (Ridge regression)

- **För många x-variabler** $\Rightarrow$ MK-metoden överanpassar data.
- Modellen är **överparametriserad**.
- Variabelselektion (t ex forward selection) är en lösning.
- **L2-regularisering** minimerar en **straffad SSE**:

$$Q_\lambda = \sum_{i=1}^{n}(y_i - a - b_1 x_{1i} - \ldots - b_k x_{ki})^2 + \lambda \cdot \sum_{j=1}^{k} b_j^2$$

- Stort λ kommer **krympa** estimaten av b_j mot noll.
- Skattningen är nu **biased**, men har **lägre varians**.
- **Bias-Variance trade-off**.

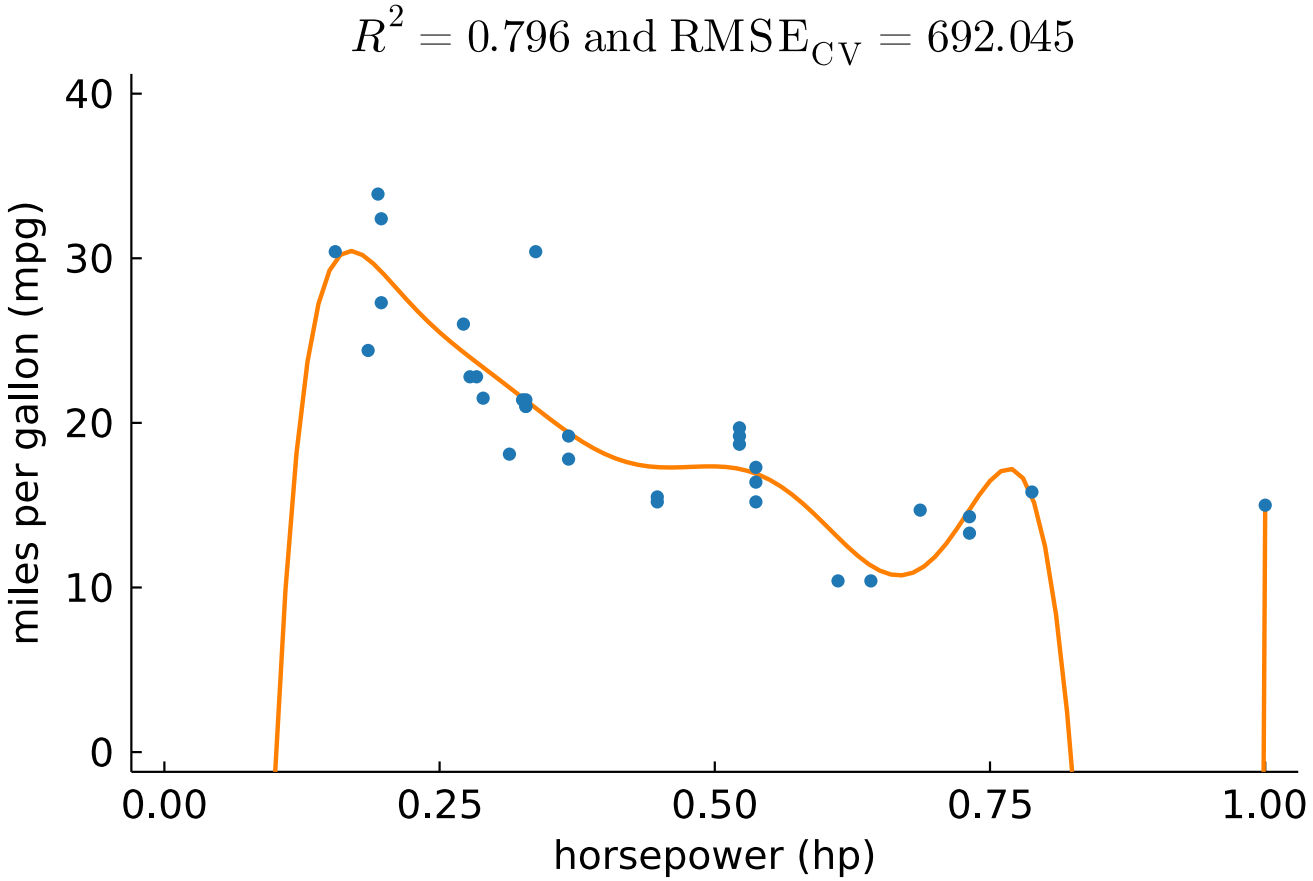
L1-regularisering (Lasso regression)

- **L1-regularisering** (**Lasso**) straffar med **absolutbelopp**:

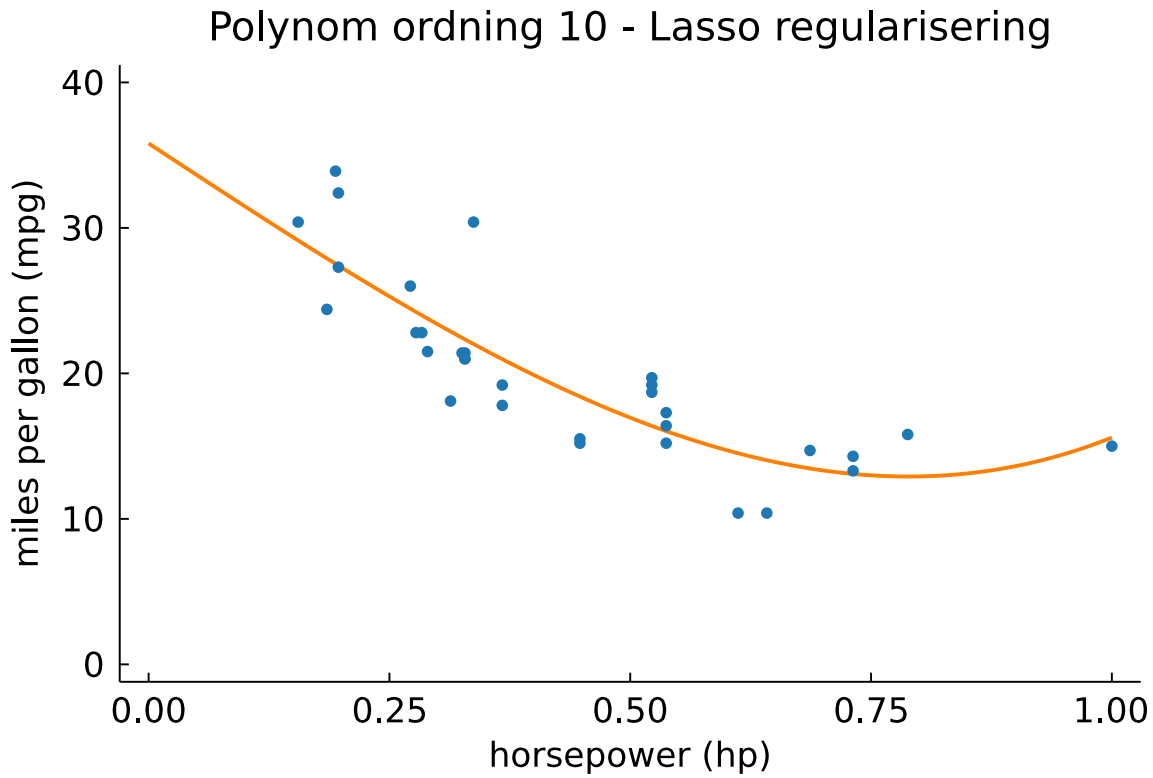
$$Q_{-} = \sum_{i=1}^{n}(y_i - a - b_1 x_{1i} - \ldots - b_k x_{ki})^2 + \lambda \cdot \sum_{j=1}^{k} |b_j|$$

- Lasso har två effekter:
  - krymper b_j mot noll (**shrinkage**)
  - kan sätta vissa b_j exakt till noll (**selection**)
- `glmnet` paketet i R gör både L1 och L2 regularisering och mer.
- Lasso är extremt populär. Go-to när man har väldigt många förklarande variabler.

Polynom ordning 10 - ingen regularisering



Polynom ordning 10 - L1-regularisering



$$Q_\lambda = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - \ldots - b_k x_{ki})^2 + \lambda \cdot \sum_{j=1}^{k} |b_j|$$

- $a = 35.80,\ b_1 = -43.58,\ b_3 = 23.32.$
- $b_2 = 0$ och $b_4 = \ldots = b_{10} = 0$ (**variabelselektion**).