Regressions- och tidsserieanalys

Föreläsning 4 - Multipel regression

Mattias Villani

Statistiska institutionen
Stockholms universitet

Institutionen för datavetenskap
Linköpings universitet


mattiasvillani.com

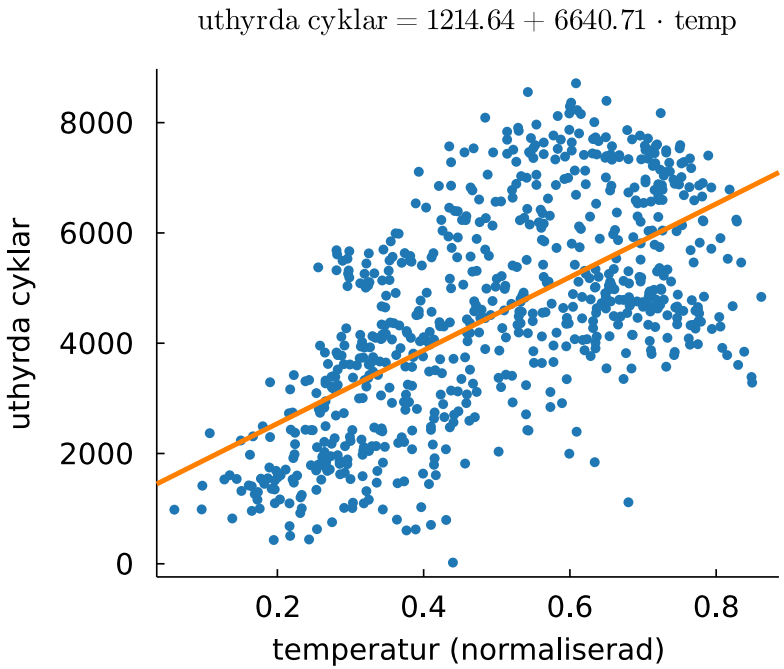

@matvil

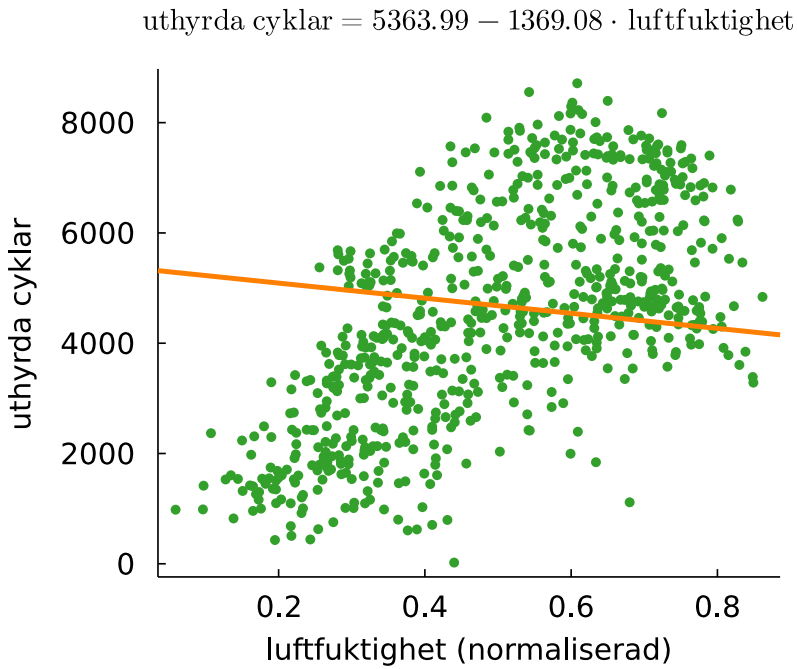

mattiasvillani

Översikt

- Multipel regression
- Hypotesttest (t cch F)
- Intro till regularisering
- Modellutvärdering

Cykeluthyrning revisited





Fler förklarande variabler - multipel regression

- Dåligt: skatta enkel regression för varje förklarande variabel.
- Bra: skatta multipel regression med alla förklarande variabler.
- Regressionanpassning med två förklarande variabler

$$y = a + b_1 x_1 + b_2 x_2$$

- b_1 talar om hur y förändras när vi ändrar x_1 med en enhet (utan att ändra x_2).
- b_2 talar om hur y förändras när vi ändrar x_2 med en enhet (utan att ändra x_1).
- I multipel regression **kontrollerar** man **för** (tar hänsyn till) de **andra förklarande variablernas effekt** på y.

Minsta kvadrat-skattningar

- **Stickprov**: (y_i, x_{1i}, x_{2i}) för $i = 1, \ldots, n$.
- x_{1i} är t ex den i:te observationens värde på x_1-variabeln.
- Hitta a, b_1 och b_2 som **minimerar residualkvadratsumman**

$$Q = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i})^2$$

- Vi får nu tre ekvationer (från partialderivatorna) som ska lösa med avseende på a, b_1 och b_2. Se AJÅ.
- Med k förklarande variabler får $k + 1$ ekvationer att lösa.

$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

- Använd dator! (enkelt programmera själv med linjär algebra).

Enkel regression temp - SAS

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1078688585	1078688585	473.47	<.0001
Error	729	1660846807	2278254		
Corrected Total	730	2739535392			

Root MSE	1509.38845	R-Square	0.3937
Dependent Mean	4504.34884	Adj R-Sq	0.3929
Coeff Var	33.50958		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1214.64212	161.16353	7.54	<.0001
temp	1	6640.71000	305.18803	21.76	<.0001

- Skattad modell

$$\text{antal uthyrningar} = 1214.64 + 6640.71 \cdot \text{temperatur}$$

- SAS-kod finns på webbsidan.

Multipel regression temp och hum - SAS

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1169231889	584615944	271.03	<.0001
Error	728	1570303503	2157010		
Corrected Total	730	2739535392			

Root MSE	1468.67638	R-Square	0.4268
Dependent Mean	4504.34884	Adj R-Sq	0.4252
Coeff Var	32.60574		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2657.89512	272.42279	9.76	<.0001
temp	1	6886.97373	299.37906	23.00	<.0001
hum	1	-2492.85413	384.76433	-6.48	<.0001

- Skattad modell:

$$\text{antal uthyrningar} = 2657.9 + 6886.97 \cdot \text{temperatur} - 2492.85 \cdot \text{luftfuktighet}$$

Multipel regression temp, hum, wind - SAS

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1262638191	420879397	207.18	<.0001
Error	727	1476897201	2031495		
Corrected Total	730	2739535392			

Root MSE	1425.30539	R-Square	0.4609
Dependent Mean	4504.34884	Adj R-Sq	0.4587
Coeff Var	31.64287		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4084.36338	337.86220	12.09	<.0001
temp	1	6625.53271	293.08535	22.61	<.0001
hum	1	-3100.12313	383.99161	-8.07	<.0001
windspeed	1	-4806.92932	708.90424	-6.78	<.0001

- Skattad modell:

$$\text{antal uthyrningar} = 4084.4 + 6625.5 \cdot \text{temp} - 3100.1 \cdot \text{hum} - 4806.9 \cdot \text{wind}$$

Multipel regression

- **Multipel** regression med k förklarande variabler:

$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

- **Residualvariansen** mäter graden av spridning kring linjen

$$s_e^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - (k+1)},$$

där de predikterade värden ges av regressionekvationen

$$\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_{ki} x_k.$$

- **Andel förklarad variation**

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

- Alternativt sätt (kom ihåg att $\text{SST} = \text{SSR} + \text{SSE}$)

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Multipel regression som sannolikhetsmodell

- **Populationsmodell** för regression med två förklarande variabler:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

- **Populationsmodell för multipel regression** med k förklarande variabler:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

- β_j talar om hur y förändras när vi ändrar x_j med en enhet (utan att ändra de andra x-variablerna).

- Samma **antaganden** som tidigare:
  - Feltermerna ε_i har **samma varians** σ_ε^2 (homoskedastiticitet)
  - Feltermerna är **normalfördelade**
  - Feltermerna är **oberoende**.

Konfidensintervall

- **Exakt 95% konfidensintervall för β_j**

$$b_j \pm t_{0.975}(n-k-1) \cdot s_{b_j}$$

där s_{b_j} är standardfelet för b_j (liknande b, men mer komplex).

- Cykeluthyrning med $k = 3$ förklarande variabler

$$t_{0.975}(n-k-1) = t_{0.975}(727) = 1.963$$

- *p*-värdet beräknas på samma sätt som i enkel regression, men från $t_{0.975}(n-k-1)$ fördelningen.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4084.36338	337.86220	12.09	<.0001
temp	1	6625.53271	293.08535	22.61	<.0001
hum	1	-3100.12313	383.99161	-8.07	<.0001
windspeed	1	-4806.92932	708.90424	-6.78	<.0001

Signifikanstest för en regressionkoefficient t-test

- **Nollhypotes** som testar om x_j är en signifikant variabel

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

- **Teststatistiska**

$$t = \left| \frac{b_j - 0}{s_{b_j}} \right|$$

- Vi förkastar nollhypotesten på signifikansnivån $\alpha = 0.05$ om

$$t_{\text{obs}} > t_{\text{crit}} = t_{0.975}(n - k - 1) \text{ (från tabell).}$$

- Cykeluthyrning. Testa om windspeed är en signifikant variabel:

$$t_{\text{obs}} = |(-4806.92 - 0)/708.90| = 6.780$$

och $t_{\text{crit}} = t_{0.975}(727) = 1.963$. Eftersom $t_{\text{obs}} > t_{\text{crit}}$ så förkastar vi H_0 på 5% signifikansnivå.

ANOVA - medelversionen

- **Mean Squared Error (MSE)**

$$\text{MSE} = \frac{\text{SSE}}{n-(k+1)} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(k+1)} = s_e^2$$

- **Mean Square Regression (MSR)**

$$\text{MSR} = \frac{\text{SSR}}{k}$$

- **Mean Square Total (MST)**

$$\text{MST} = \frac{\text{SST}}{n-1}$$

- Notera att **frihetsgraderna summerar också**

$$\begin{aligned} \text{df}(\text{SST}) &= \text{df}(\text{SSE}) + \text{df}(\text{SSR}) \\ n-1 &= n-(k+1) + k \end{aligned}$$

Signifikanstest för flera regressionkoefficienter

- **F-test statistiska**

$$F = \frac{\text{MSR}}{\text{MSE}}$$

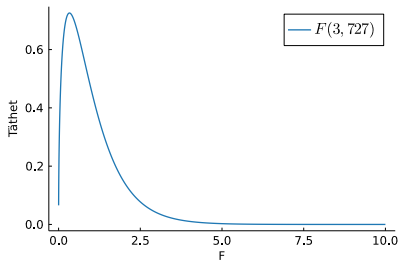
- **Nollhypotesen** om **ingen regression**

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
$$H_1 : \text{åtminstone något } \beta_j \neq 0$$

- Under H_0 följer F en **F-fördelning med k och $n-(k+1)$ frihetsgrader**.

$$F \sim F(k, n-k-1)$$

- Cykeluthyrningsdata: $F_{\text{obs}} = 207.18$. $F_{0.95}(3, 727) = 2.617$. Vi tokförkastar nollhypotesen om ingen regression!



Val av förklarande variabler

- Ju fler förklarande variabler desto mer förklarar regressionen.
- R^2 **kan inte minska** när man lägger till fler förklarande variabler. Se upp för överanpassning!
- R^2_{adjusted} ("justerad R-2"), se AJÅ, kan minska om en förklarande variabel bara reducerar variationen marginellt.
- Andra vanliga informationskriterier: **AIC**, **BIC**.
- **Full sökning**: Gå igenom alla möjliga kombinationer av förklarande variabler och välj modell med högst R^2_{adjusted}. Beräkningstungt.

Stepwise selection and beyond

- **Forward selection**:
  1. Börja med bara interceptet.
  2. Lägg till x-variabeln med högst t_{obs}, om $t_{\text{obs}} > 2$, annars stanna.
  3. Lägg till x-variabeln med högst t_{obs}, givet att valda variabeln i Steg 2 ingår i modellen, om $t_{\text{obs}} > 2$, annars stanna.
  4. Fortsätt tills ingen ny förklarande variabel har $t_{\text{obs}} > 2$ i modellen där alla tidigare variabler ingår.
- **Backward selection**. Starta med alla variabler i modellen. Ta bort den variabel som har lägst t_{obs}. Skatta modellen utan denna variabel. Fortsätt tills alla variabler som är kvar har $t_{\text{obs}} > 2$.
- Det finns massor av andra (bättre) variabelselektionsstrategier. **Bayesian variable selection**. *Bayesian Learning 7.5 hp*.

L2-regularisering (Ridge regression)

- **För många förklarande variabler** $\Rightarrow$ MK-metoden överanpassar data. Modellen är **överparametriserad**.
- Variabelselektion försöker minska antalet skattade parametrar.
- **L2-regularisering** (**ridge regression**) behåller alla variabler i modellen men minimerar en **straffad residualkvadratsumma**:

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

- Straff/kostnad för att introducera en variabel i modellen

$$\lambda \cdot \sum_{j=1}^{k} b_j^2$$

- Hur hårt vi straffar bestäms av **regulariseringsparametern** λ.
- Stort λ kommer krympa estimaten av b_j mot noll. Biased, men lägre varians. **Bias-Variance trade-off**.
- Vi kan bestämma λ själva, eller skatta via korsvalidering.

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)
- SAS GLMSELECT med SELECTION=LASSO som option gör Lasso regression.
- glmnet paketet i R gör Lasso och mycket mer.
- Lasso är extremt populär. Go-to när man har väldigt många förklarande variabler.

Prognosförmåga på testdata

- Välj den modell som ger bäst prediktioner på nya (test) data.
- Dela upp observationer i två delmängder:
  - **Träningsdata** för att skatta modellens parametrar.
  - **Testdata** för att utvärdera modellens prediktioner.
- Modellen får aldrig chans att anpassa sig till testdata.
- Prediktionsmått: **kvadrerade prediktionsfel på testdata**

$$Q_{\text{test}} = \sum_{j=1}^{n_{\text{test}}} (y_j - \hat{y}_j)^2$$

- Observera:
  - summan är över observationerna i testdata.
  - modellen som ger $\hat{y}_j$ är skattad enbart på träningsdata.
  - överanpassning på träningsdata $\Rightarrow$ dåliga prediktioner på testdata.

Korsvalidering

- Vilka observationer ska vara i träning respektive test? **Korsvalidering**.
- Mått på modellens prognosförmåga: genomsnittligt Q_{test} över alla $K = 3$ testdataset.

Split 1

country	spending (x)	lifespan (y)
Australia	3.357	81.4
Austria	3.763	80.1
Belgium	3.595	79.8
Canada	3.895	80.7
Czech	1.626	77
Denmark	3.512	78.4
Finland	2.84	79.5
France	3.601	81
Germany	3.588	80
Greece	2.727	79.5
Hungary	1.388	73.3
Iceland	3.319	81.2
Ireland	3.424	79.7
Italy	2.686	81.4
Japan	2.581	82.6
Korea	1.688	79.4
Luxembourg	4.162	79.4
Mexico	0.823	75
Netherlands	3.837	80.2
N.Zealand	2.454	80.2
Norway	4.763	80.6
Poland	1.035	75.4
Portugal	2.15	79.1
Slovakia	1.555	74.3
Spain	2.671	81
Sweden	3.323	81
Switzerland	4.417	81.9
Turkey	0.618	73.4
UK	2.992	79.5
USA	7.29	78.1

Split 2

country	spending (x)	lifespan (y)
Australia	3.357	81.4
Austria	3.763	80.1
Belgium	3.595	79.8
Canada	3.895	80.7
Czech	1.626	77
Denmark	3.512	78.4
Finland	2.84	79.5
France	3.601	81
Germany	3.588	80
Greece	2.727	79.5
Hungary	1.388	73.3
Iceland	3.319	81.2
Ireland	3.424	79.7
Italy	2.686	81.4
Japan	2.581	82.6
Korea	1.688	79.4
Luxembourg	4.162	79.4
Mexico	0.823	75
Netherlands	3.837	80.2
N.Zealand	2.454	80.2
Norway	4.763	80.6
Poland	1.035	75.4
Portugal	2.15	79.1
Slovakia	1.555	74.3
Spain	2.671	81
Sweden	3.323	81
Switzerland	4.417	81.9
Turkey	0.618	73.4
UK	2.992	79.5
USA	7.29	78.1

Split 3

country	spending (x)	lifespan (y)
Australia	3.357	81.4
Austria	3.763	80.1
Belgium	3.595	79.8
Canada	3.895	80.7
Czech	1.626	77
Denmark	3.512	78.4
Finland	2.84	79.5
France	3.601	81
Germany	3.588	80
Greece	2.727	79.5
Hungary	1.388	73.3
Iceland	3.319	81.2
Ireland	3.424	79.7
Italy	2.686	81.4
Japan	2.581	82.6
Korea	1.688	79.4
Luxembourg	4.162	79.4
Mexico	0.823	75
Netherlands	3.837	80.2
N.Zealand	2.454	80.2
Norway	4.763	80.6
Poland	1.035	75.4
Portugal	2.15	79.1
Slovakia	1.555	74.3
Spain	2.671	81
Sweden	3.323	81
Switzerland	4.417	81.9
Turkey	0.618	73.4
UK	2.992	79.5
USA	7.29	78.1

Träning
Test

- *Maskininlärning 7.5 hp.*