

Regressions- och tidsserieanalys

Föreläsning 4 - Multipel regression

Mattias Villani

Statistiska institutionen
Stockholms universitet

Institutionen för datavetenskap
Linköpings universitet



mattiasvillani.com



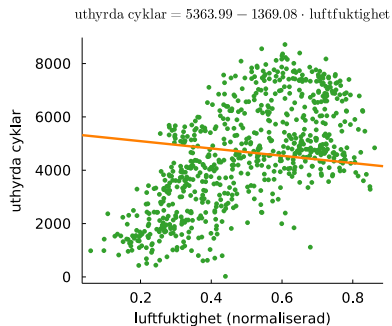
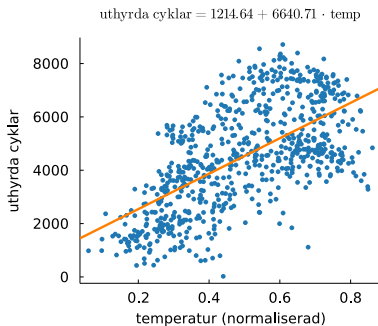
@matvil



mattiasvillani

- Multipel regression
- Hypotesttest (t och F)
- Variabelselektion

Cykeluthyrning revisited



Fler förklarande variabler - multipel regression

- Skatta enkel regression för varje förklarande variabel. 🙅
- Skatta multipel regression med alla förklarande variabler. 👍
- Regressionanpassning med två förklarande variabler

$$y = a + b_1x_1 + b_2x_2$$

- 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 variabelernas effekt** på y .

Minsta kvadrat-skattningar

- **Stickprov:** (y_i, x_{1i}, x_{2i}) för $i = 1, \dots, n$.
- 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 lösas med avseende på a , b_1 och b_2 . Se AJÅ.
- Med k förklarande variabler får vi $k + 1$ ekvationer att lösa.

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

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

Enkel regression temp - R

```
> lmfit = lm(nRides ~ temp, data = bike); regsummary(lmfit)
```

Analysis of variance - ANOVA

```
-----
```

	df	SS	MS	F	Pr(>F)
Regr	1	1078688585	1078688585	473.47	2.8106e-81
Error	729	1660846807	2278254		
Total	730	2739535392			

Measures of model fit

```
-----
```

Root MSE	R2	R2-adj
1509.38845	0.39375	0.39292

Parameter estimates

```
-----
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1214.6	161.16	7.5367	1.4327e-13
temp	6640.7	305.19	21.7594	2.8106e-81

■ Skattad modell

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

Multipel regression temp och hum - R

```
> lmfit = lm(nRides ~ temp + hum, data = bike); regsummary(lmfit)
```

Analysis of variance - ANOVA

```
-----  
              df          SS          MS          F          Pr(>F)  
Regr         2 1169231889 584615944 271.03 1.0559e-88  
Error       728 1570303503   2157010  
Total       730 2739535392
```

Measures of model fit

```
-----  
      Root MSE          R2          R2-adj  
1468.67638      0.42680      0.42522
```

Parameter estimates

```
-----  
              Estimate Std. Error t value  Pr(>|t|)  
(Intercept)   2657.9      272.42  9.7565 3.2258e-21  
temp          6887.0      299.38 23.0042 1.9558e-88  
hum           -2492.9      384.76 -6.4789 1.7012e-10
```

■ Skattad modell:

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

Multipel regression temp, hum, wind - R

```
> lmfit = lm(nRides ~ temp + hum + windspeed, data = bike); regsummary(lmfit)
```

Analysis of variance - ANOVA

```
-----  
              df          SS          MS          F          Pr(>F)  
Regr         3 1262638191 420879397 207.18 4.2551e-97  
Error      727 1476897201   2031495  
Total      730 2739535392
```

Measures of model fit

```
-----  
      Root MSE          R2          R2-adj  
1425.30539      0.46090      0.45867
```

Parameter estimates

```
-----  
              Estimate Std. Error t value  Pr(>|t|)  
(Intercept)   4084.4      337.86 12.0888 8.7098e-31  
temp          6625.5      293.09 22.6062 4.1807e-86  
hum           -3100.1      383.99 -8.0734 2.8330e-15  
windspeed     -4806.9      708.90 -6.7808 2.4754e-11
```

■ Skattad modell:

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:

$$y = a + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

- **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

$$\hat{y}_i = a + b_1x_{1i} + b_2x_{2i} + \dots + b_kx_{ki}.$$

- **Andel förklarad variation**

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

- Alternativt sätt (kom ihåg att $SST = SSR + SSE$)

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

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:

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

- 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 (homoskedasticitet)
 - ▶ 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

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

Parameter estimates

	Estimate	Std. Error	t value	Pr(> t)	2.5 %	97.5 %
(Intercept)	4084.4	337.86	12.0888	8.7098e-31	3421.1	4747.7
temp	6625.5	293.09	22.6062	4.1807e-86	6050.1	7200.9
hum	-3100.1	383.99	-8.0734	2.8330e-15	-3854.0	-2346.3
windspeed	-4806.9	708.90	-6.7808	2.4754e-11	-6198.7	-3415.2

```
> lmfit = lm(nRides ~ temp + hum + windspeed, data = bike);  
> regsummary(lmfit, conf_intervals = T)
```

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

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

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

Signifikanstest för flera regressionkoefficienter

■ F-test statistiska

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

■ Nollhypotesen om ingen regression

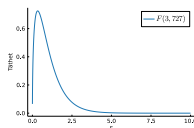
$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{åtminstone något } \beta_j \neq 0$$

■ 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.*