

Regressions- och tidsserieanalys

Föreläsning 5 - Modeller: antaganden, kontroll och utvärdering

Mattias Villani

Statistiska institutionen
Stockholms universitet

Institutionen för datavetenskap
Linköpings universitet



mattiasvillani.com



@matvil



mattiasvillani

- Modellkontroll
- Binära och kategoriska förklarande variabler.
- Modellutvärdering

Multipel linjär regression - antaganden

■ Populationsmodell för multipel regression:

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

■ Antaganden:

- ▶ Betingade väntevärdet $\mu_{y|x}$ är en **linjär funktion** av x
- ▶ Feltermerna ε_i har **samma varians** σ_ε^2 (homoskedasticitet)
- ▶ Feltermerna är **normalfördelade**
- ▶ Feltermerna är **oberoende**.

Antagandet om linjäritet, normalitet och oberoende

■ Linjäritet:

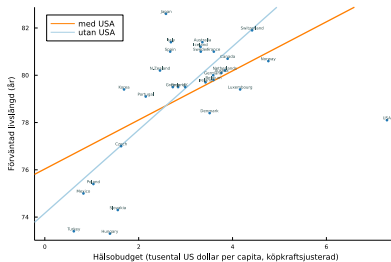
- ▶ **Plotta residualerna** mot varje förklarande variabel.
- ▶ **Testa om icke-linjära effekter** är signifikanta (se F7).

■ Normalitet:

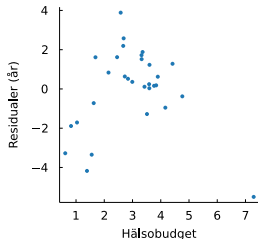
- ▶ **Histogram** över **residualerna**
- ▶ **Q-Q-plot** för **residualerna**
- ▶ **Normalitetstest**

- **Oberoende residualer?** Ofta problem när variabler i regression är **observerade över tid**. Ex. cykeluthyrningsdata. Återkommer till detta när vi pratar om tidsserier.

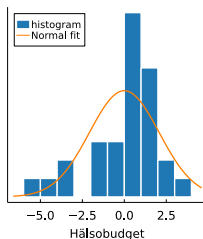
Hälsobudgetdata med USA



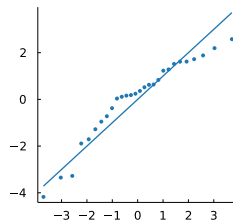
Residualer vs hälsobudget



Histogram residualer

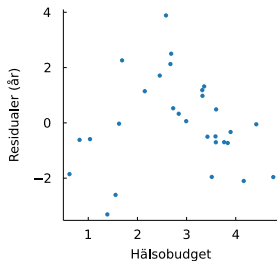


QQ-plot

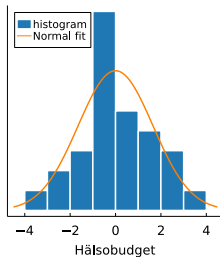


Hälsobudgetdata - utan USA

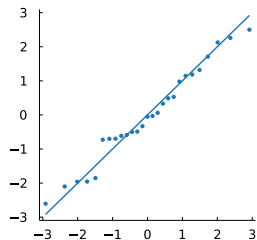
Residualer vs hälsobudget



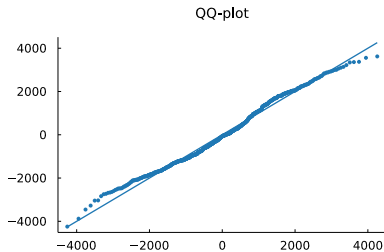
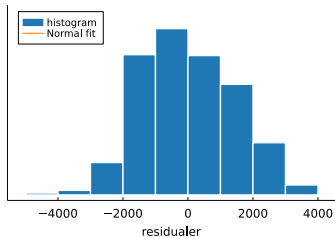
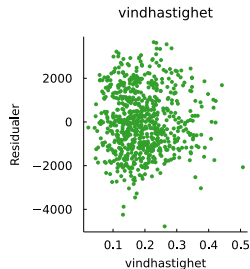
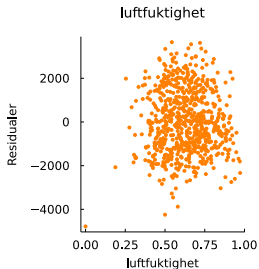
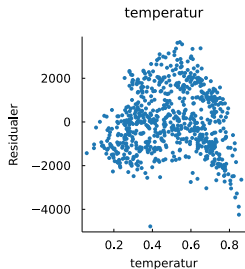
Histogram residualer



QQ-plot



Cykeluthyrningar



Antagandet om konstant varians

- Plotta residualerna mot varje förklarande variabel.
- Test för heteroskedasticitet:

H_0 : feltermerna har samma varians (homoskedastiska)

H_1 : feltermerna har olika varians (heteroskedastiska)

- Testprocedur:

- ▶ skatta regression med **kvadrerade residualer e^2 som y-variabel**

$$e^2 = \tilde{\alpha} + \tilde{\beta}_1 x_1 + \dots + \tilde{\beta}_k x_k + \tilde{\varepsilon}$$

- ▶ använd t ex F -test för att testa $H_0 : \tilde{\beta}_1 = \dots = \tilde{\beta}_k = 0$.
- ▶ om F -testet förkastas så förkastar vi homoskedasticitet.

- AJÅ: kvadrater x_1^2, \dots, x_k^2 som förklarande variabler i regressionen för e^2 . Kollar om variansen är ett **icke-linjär funktion** av någon förklarande variabel. Se F7.

Multikollinearitet

- Förklarande variabler är ofta **korrelerade**.
- **Multikollinearitet** - linjära beroenden mellan olika x_j .

```
Title "korrelationer";  
proc corr data = work.cykeluthyr;  
var temp hum windspeed;  
run;
```

Pearson Correlation Coefficients, N = 731 Prob > r under H0: Rho=0			
	temp	hum	windspeed
temp	1.00000	0.12696 0.0006	-0.15794 <.0001
hum	0.12696 0.0006	1.00000	-0.24849 <.0001
windspeed	-0.15794 <.0001	-0.24849 <.0001	1.00000

- **Problem vid multikollinearitet:**
 - ▶ **svårt att separera** de olika förklarande variabelernas effekt på y
 - ▶ **stora standardfel** för b_j .
 - ▶ **insignifikans**.
- **Prediktioner påverkas inte** av multikollinearitet.

Variance inflation factors

- **Variance Inflation Factor (VIF)** för förklarande variabeln x_j

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

- R_j^2 är förklaringsgraden i regressionen **med x_j som responsvariabel** och alla andra x som förklarande variabler.
- Tumregel: $\text{VIF} > 10$ är stark multikollinearitet.
- Cykeluthyrning. Ny variabel: *upplevd temperatur* (feelttemp).

variable	R^2	VIF	variable	R^2	VIF
temp	0.033	1.034	temp	0.984	62.969
hum	0.070	1.075	feelttemp	0.984	63.632
windspeed	0.078	1.085	hum	0.073	1.079
			windspeed	0.113	1.127

Binära förklarande variabler

- Binära (dummy) variabler som bara kan anta två värden. Ex:

$$\text{holiday} = \begin{cases} 1 & \text{om röd dag} \\ 0 & \text{annars} \end{cases}$$

$$\text{workingday} = \begin{cases} 1 & \text{om arbetsdag} \\ 0 & \text{om helg eller arbetsfri dag} \end{cases}$$

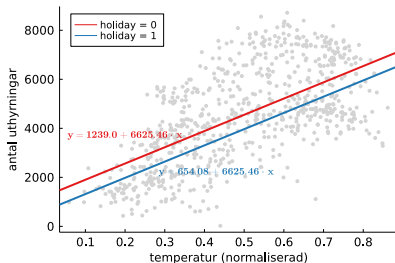
- Varianter av kodning: (0,1) eller (-1,1), eller (true,false).
- Regressionsmodell med binär förklarande variabel:

$$y = \alpha + \beta_1 \cdot \text{temp} + \beta_2 \cdot \text{workingday} + \varepsilon$$

innebär att vi får två parallella regressionlinjer

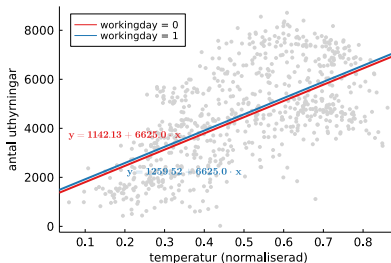
$$y = \begin{cases} \alpha + \beta_1 \cdot \text{temp} + \varepsilon & \text{om workingday} = 0 \\ (\alpha + \beta_2) + \beta_1 \cdot \text{temp} + \varepsilon & \text{om workingday} = 1 \end{cases}$$

Binära förklarande variabler



Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits
Intercept	1	1239.00123	161.53482	7.67	<.0001	921.87157 1556.13090
temp	1	6625.45780	304.88015	21.73	<.0001	6026.90857 7224.00704
holiday	1	-584.91862	333.87396	-1.75	0.0802	-1240.38930 70.55207

```
proc reg data = work.cykeluthyr;  
model NRIDES = TEMP HOLIDAY / clb alpha=0.05;  
run;
```



Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits
Intercept	1	1142.13473	177.46139	6.44	<.0001	793.73757 1490.53190
temp	1	6624.99960	305.62194	21.68	<.0001	6024.99407 7225.00512
workingday	1	117.38410	120.25018	0.98	0.3293	-118.69442 353.46262

```
proc reg data = work.cykeluthyr;  
model NRIDES = TEMP WORKINGDAY / clb alpha=0.05;  
run;
```

Ännu fler binära förklarande variabler

Multipel linjär regression - nRides mot temp och hum

The REG Procedure
Model: MODEL1
Dependent Variable: nRides

Number of Observations Read	731
Number of Observations Used	731

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1995735713	332622619	323.77	<.0001
Error	724	743799679	1027348		
Corrected Total	730	2739535392			

Root MSE	1013.58158	R-Square	0.7285
Dependent Mean	4504.34884	Adj R-Sq	0.7262
Coeff Var	22.50229		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	1	2577.86429	252.55993	10.21	<.0001	2082.02702	3073.70157
temp	1	6280.85581	209.04363	30.05	<.0001	5870.45174	6691.25988
hum	1	-2220.63420	275.17032	-8.07	<.0001	-2760.86122	-1680.40718
windspeed	1	-4363.74853	504.41242	-8.65	<.0001	-5354.03420	-3373.46287
workingday	1	76.55228	83.45154	0.92	0.3593	-87.28362	240.38817
holiday	1	-607.03615	232.02130	-2.62	0.0091	-1062.55104	-151.52126
yr	1	2008.60930	75.63209	26.56	<.0001	1860.12490	2157.09370

Kategoriska förklarande variabler

- Kategoriska (klass) förklarande **variabler**. Ex:

$$\text{season} = \begin{cases} 1 & \text{om vinter} \\ 2 & \text{om vår} \\ 3 & \text{om sommar} \\ 4 & \text{om höst} \end{cases}$$

- Koda som **fyra binära variabler**

	vinter	vår	sommar	höst	temp	...
2011-01-01	1	0	0	0	0.344	
2011-01-02	1	0	0	0	0.363	
⋮						
2011-04-28	0	1	0	0	0.453	
⋮						
2011-07-14	0	0	1	0	0.830	
⋮						
2011-10-04	0	0	0	1	0.521	

Kategoriska förklarande variabler

- Regressionen kan inte skattas pga **perfekt multikollinearitet!**

$$y = a + b_1 \cdot \text{temp} + b_2 \cdot \text{vinter} + b_3 \cdot \text{vår} + b_4 \cdot \text{sommar} + b_5 \cdot \text{höst}$$

- Lösning: **ta bort en** av de fyra dummyvariabler, t ex vinter:

$$y = a + b_1 \cdot \text{temp} + b_3 \cdot \text{vår} + b_4 \cdot \text{sommar} + b_5 \cdot \text{höst}$$

- Vinter blir nu **referenskategorin** (alla tre dummies är noll då).

- Vinterdag:

$$y = a + b_1 \cdot \text{temp}$$

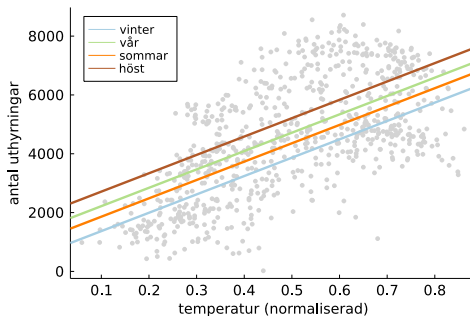
- Vårdag:

$$y = (a + b_3) + b_1 \cdot \text{temp}$$

- Koefficienten b_3 är **hur många fler** cyklar hyrs ut under en vårdag **jämfört med en vinterdag**.

- Koefficienten b_4 är hur många fler cyklar hyrs ut under en sommardag **jämfört med en vinterdag**.

Cykeluthyrning - säsongsdummies



$nRides \sim 1 + temp + spring + summer + fall$

Coefficients:

	Coef.	Std. Error	t	Pr(> t)	Lower 95%	Upper 95%
(Intercept)	745.787	187.476	3.98	<1e-04	377.728	1113.85
temp	6241.35	518.142	12.05	<1e-29	5224.11	7258.58
spring	848.724	197.082	4.31	<1e-04	461.806	1235.64
summer	490.196	259.006	1.89	0.0588	-18.2936	998.685
fall	1342.87	164.588	8.16	<1e-14	1019.75	1666.0

F-test för en grupp av förklarande variabler

- **Testa om det finns en säsongseffekt?** Vi kan t -testa varje säsongsdummy (vår, sommar, höst).
- **F -test** kan användas för att **testa en grupp av variabler**

$$H_0 : \beta_{\text{vår}} = \beta_{\text{sommar}} = \beta_{\text{höst}} = 0$$

H_1 : någon av $\beta_{\text{vår}}, \beta_{\text{sommar}}$ eller $\beta_{\text{höst}}$ är skild från noll.

- **Teststatistiska**

$$F = \frac{(R_{\text{UR}}^2 - R_{\text{R}}^2) / r}{(1 - R_{\text{UR}}^2) / (n - k - 1)}$$

- ▶ R_{UR}^2 är R^2 för regressionen **U**tan nollhypotesens **R**estriktioner (de tre säsongsdummies är med i modellen)
 - ▶ R_{R}^2 är R^2 för regressionen med nollhypotesens **R**estriktioner (de tre säsongsdummies är inte med i modellen)
 - ▶ r är antalet restriktioner under H_0 , dvs $r = 3$ här.
- Under H_0 följer teststatistikan F en $F(r, n - k - 1)$ -fördelning.

F-test för säsong i cykeluthyrningsdata

- Under H_0 : temp, hum, windspeed.
- Under H_1 : temp, hum, windspeed, vår, sommar, höst.
- Så $k = 6$ och $r = 3$.
- $R_{UR}^2 = 0.5354$
- $R_R^2 = 0.4609$

$$F_{\text{obs}} = \frac{(0.5354 - 0.4609)/3}{(1 - 0.5354) / (731 - 6 - 1)} = 38.698$$

$$F_{\text{crit}} = F_{0.95}(3, 724) = 2.617$$

- $F_{\text{obs}} > F_{\text{crit}}$ så nollhypotesen förkastas på signifikansnivån 5%.
Det verkar finnas en säsongseffekt.