

# Regressions- och tidsserieanalys

## Föreläsning 5 - Modeller: antaganden, kontroll och utvärdering

**Mattias Villani**

Statistiska institutionen  
Stockholms universitet

Institutionen för datavetenskap  
Linköpings universitet



[mattiasvillani.com](http://mattiasvillani.com)



[@matvil](https://twitter.com/matvil)



[mattiasvillani](https://github.com/mattiasvillani)

- Modellkontroll
- Binära och kategoriska förklarande variabler.
- Modellutvärdering

# Multipl linjär regression - antaganden

## ■ Populationsmodell för multipl 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:

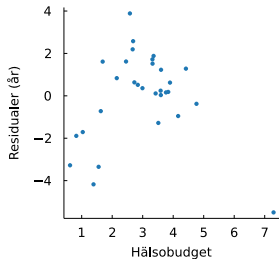
- ▶ Betingade väntevärdet  $\mu_{y|x}$  är en **linjär funktion** av  $x$
- ▶ Feltermerna  $\varepsilon_i$  har **samma varians**  $\sigma_\varepsilon^2$  (homoskedasticitet)
- ▶ Feltermerna är **normalfördelade**
- ▶ Feltermerna är **oberoende**.

# Antagandet om linjäritet, normalitet och oberoende

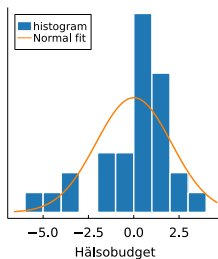
- **Linjäritet:** Plotta residualerna mot varje förklarande variabel.
- **Normalitet:**
  - ▶ **Histogram** över **residualerna**
  - ▶ **Q-Q-plot** för **residualerna**
  - ▶ **Normalitetstest**
- **Oberoende residualer?** Ofta problem när variabler i regression är **observerade över tid**. Ex. cykeluthyrningsdata. Återkommer till detta när vi pratar om tidsserier.

# Hälsobudgetdata

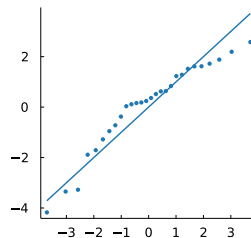
Residualer vs hälsobudget



Histogram residualer

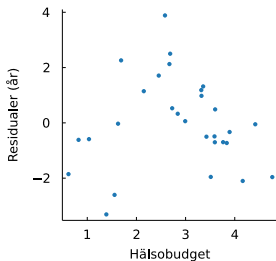


QQ-plot

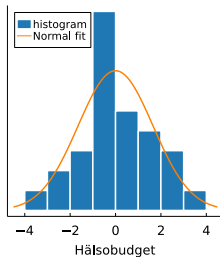


# Hälsobudgetdata - utan USA

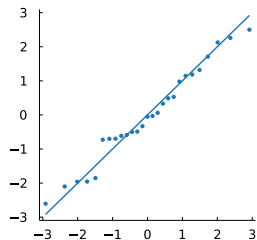
Residualer vs hälsobudget



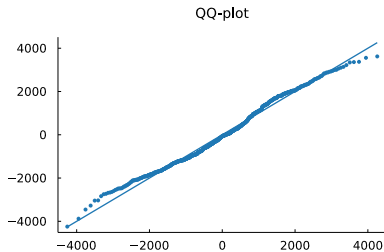
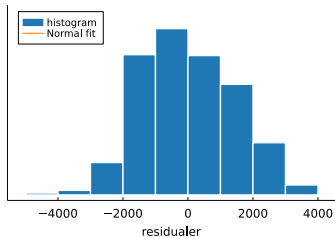
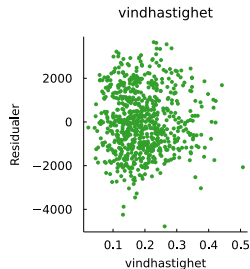
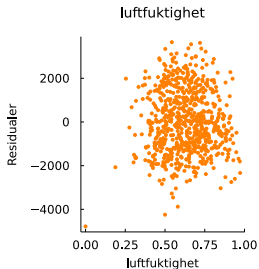
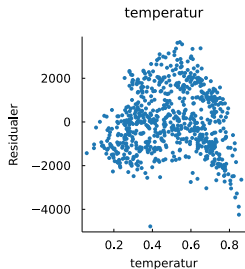
Histogram residualer



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**

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

- ▶ använd t ex  $F$ -test för att testa  $H_0 : \tilde{b}_1 = \dots = \tilde{b}_k = 0$ .
- ▶ 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.
- Cykeluthyrningsdata. Problem om vi lägger till upplevd temperatur (feeltemp) som ytterligare förklarande variabel:

variable	$R^2$	VIF	variable	$R^2$	VIF
temp	0.033	1.034	temp	0.984	62.969
hum	0.070	1.075	feeltemp	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: cykeluthyrningsdata:

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

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

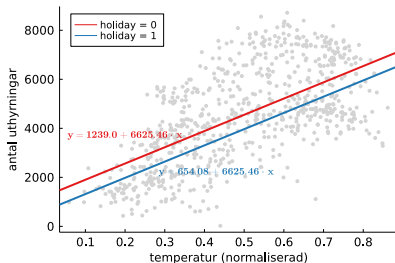
- Kan kodas på olika sätt, t ex  $-1$  och  $1$ .
- Regressionsmodell med binära förklarande variabler

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

innebär att vi får **två parallella regressionlinjer**

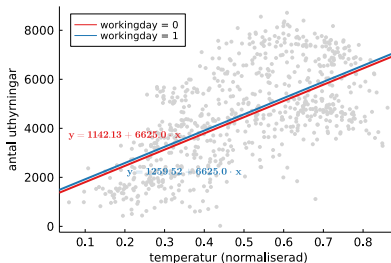
$$y = \begin{cases} \alpha + \beta_1 \cdot \text{temp} + \beta_2 \cdot \text{workingday} + \varepsilon & \text{om workingday} = 0 \\ (\alpha + \beta_2) + \beta_1 \cdot \text{temp} + \beta_2 \cdot \text{workingday} + \varepsilon & \text{om workingday} = 1 \end{cases}$$

# 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;
```

# Kategoriska förklarande variabler

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

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

- 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	
2011-07-15	0	0	1	0	0.780	

# Kategoriska förklarande variabler

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

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

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

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

- Vinter blir nu **referenskategorin** (all tre dummies är noll då).
- Vinterdag:

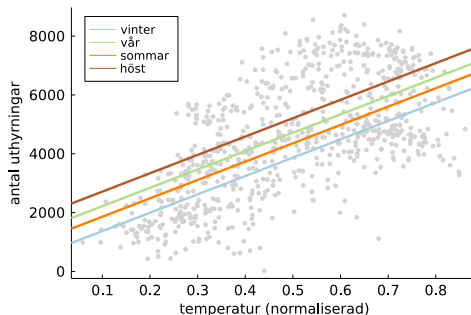
$$y = a + b_1 \cdot \text{temp}$$

- 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

# 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  $Q_{\text{test}}$  över alla  $K = 3$  testdataset.

Split 1			Split 2			Split 3		
country	spending (x)	lifespan (y)	country	spending (x)	lifespan (y)	country	spending (x)	lifespan (y)
Australia	3.357	81.4	Australia	3.357	81.4	Australia	3.357	81.4
Austria	3.763	80.1	Austria	3.763	80.1	Austria	3.763	80.1
Belgium	3.595	79.8	Belgium	3.595	79.8	Belgium	3.595	79.8
Canada	3.895	80.7	Canada	3.895	80.7	Canada	3.895	80.7
Czech	1.636	77	Czech	1.636	77	Czech	1.636	77
Denmark	3.512	78.4	Denmark	3.512	78.4	Denmark	3.512	78.4
Finland	2.84	79.5	Finland	2.84	79.5	Finland	2.84	79.5
France	3.601	81	France	3.601	81	France	3.601	81
Germany	3.588	80	Germany	3.588	80	Germany	3.588	80
Greece	2.727	79.5	Greece	2.727	79.5	Greece	2.727	79.5
Hungary	1.388	73.3	Hungary	1.388	73.3	Hungary	1.388	73.3
Iceland	3.319	81.2	Iceland	3.319	81.2	Iceland	3.319	81.2
Ireland	3.424	79.7	Ireland	3.424	79.7	Ireland	3.424	79.7
Italy	2.886	81.4	Italy	2.886	81.4	Italy	2.886	81.4
Japan	2.581	82.6	Japan	2.581	82.6	Japan	2.581	82.6
Korea	1.688	79.4	Korea	1.688	79.4	Korea	1.688	79.4
Luxembourg	4.162	79.4	Luxembourg	4.162	79.4	Luxembourg	4.162	79.4
Mexico	0.823	75	Mexico	0.823	75	Mexico	0.823	75
Netherlands	3.837	80.2	Netherlands	3.837	80.2	Netherlands	3.837	80.2
N.Zealand	2.454	80.2	N.Zealand	2.454	80.2	N.Zealand	2.454	80.2
Norway	4.763	80.6	Norway	4.763	80.6	Norway	4.763	80.6
Poland	1.035	75.4	Poland	1.035	75.4	Poland	1.035	75.4
Portugal	2.15	79.1	Portugal	2.15	79.1	Portugal	2.15	79.1
Slovakia	1.555	74.3	Slovakia	1.555	74.3	Slovakia	1.555	74.3
Spain	2.671	81	Spain	2.671	81	Spain	2.671	81
Sweden	3.323	81	Sweden	3.323	81	Sweden	3.323	81
Switzerland	4.417	81.9	Switzerland	4.417	81.9	Switzerland	4.417	81.9
Turkey	0.618	73.4	Turkey	0.618	73.4	Turkey	0.618	73.4
UK	2.992	79.5	UK	2.992	79.5	UK	2.992	79.5
USA	7.29	78.1	USA	7.29	78.1	USA	7.29	78.1
Training								
Test								

- Maskininlärning 7.5 hp.