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

Översikt

- Modellkontroll
- Binära och kategoriska förklarande variabler.

Multipel linjär regression - antaganden

- **Populationsmodell för multipel regression**:

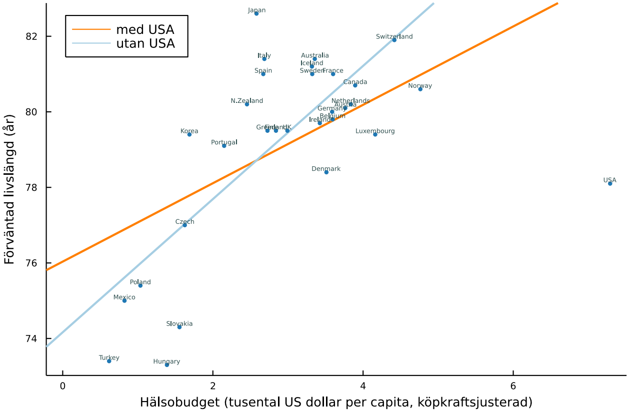
$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

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

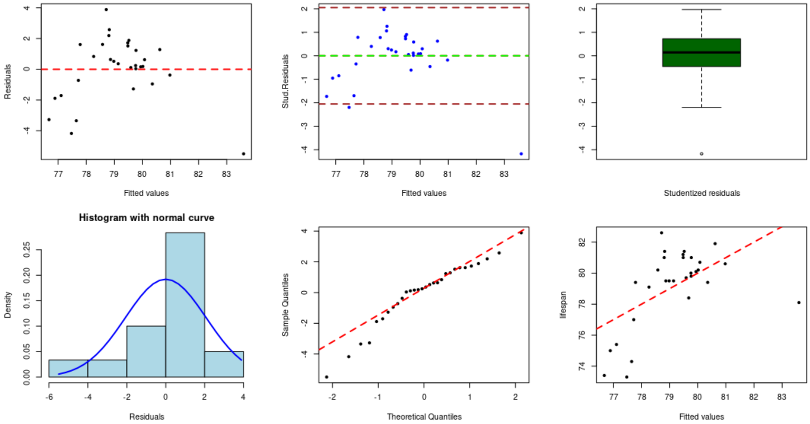
Antagandet om linjäritet, normalitet och oberoende

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

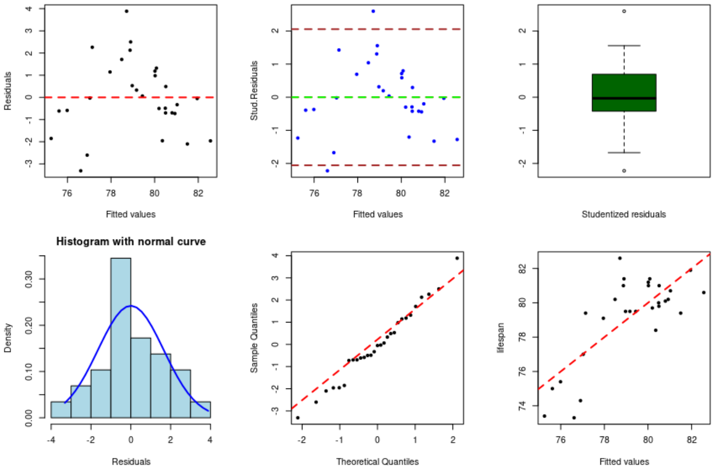
Hälsobudgetdata med USA



Hälsobudgetdata med USA

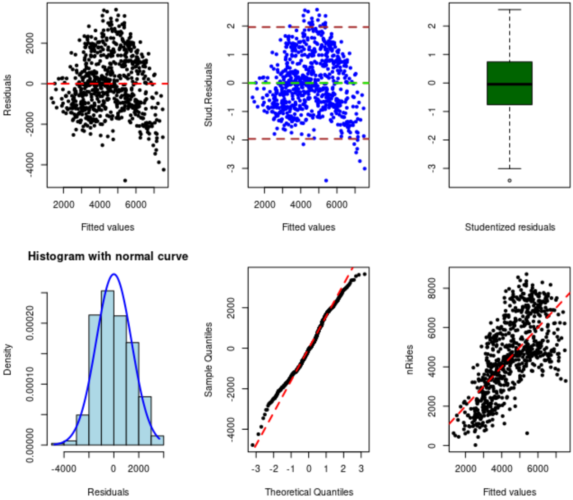


Hälsobudgetdata - utan USA



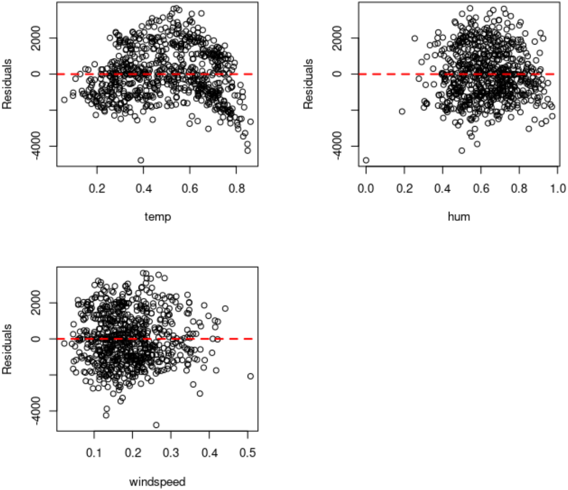
Cykeluthyrningar - residualanalys 1

- `lmfit = lm(nRides ~ temp + hum + windspeed, data = bike)`
- `res.diagnostics(lmfit)`



Cykeluthyrningar - residualanalys 2

- `lmfit = lm(nRides ~ temp + hum + windspeed, data = bike)`
- `res.diagnostics(lmfit, regressors = T)`



Antagandet om konstant varians

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

$$H_0: \text{ feltermerna har samma varians (homoskedastiska)}$$
$$H_1: \text{ feltermerna har olika varians (heteroskedastiska)}$$

- **Testprocedur**:
  - skatta regression med **kvadrerade residualer e^2 som y-variabel**
  $$e^2 = \tilde{\alpha} + \tilde{\beta}_1 x_1 + \ldots + \tilde{\beta}_k x_k + \tilde{\varepsilon}$$
  - använd t ex F-test för att testa $H_0 : \tilde{\beta}_1 = \ldots = \tilde{\beta}_k = 0$.
  - om F-testet förkastas så förkastar vi homoskedastiticitet.

- AJÅ: kvadrater $x_1^2, \ldots, x_k^2$ 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.

```
> library(Hmisc)
> X = as.matrix(bike[,c("temp","hum","windspeed")])
> rcorr(X)
           temp   hum windspeed
temp       1.00  0.13     -0.16
hum        0.13  1.00     -0.25
windspeed -0.16 -0.25      1.00

n= 731


P
          temp   hum    windspeed
temp             6e-04  0e+00
hum       6e-04         0e+00
windspeed 0e+00  0e+00
```

- **Problem vid multikollinearitet**:
  - **svårt att separera** de olika förklarande variablernas 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

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

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

variable	R^2	VIF
`temp`	0.033	1.034
`hum`	0.070	1.075
`windspeed`	0.078	1.085

variable	R^2	VIF
temp	0.984	62.969
feeltemp	0.984	63.632
hum	0.073	1.079
windspeed	0.113	1.127

Binära förklarande variabler

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

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

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

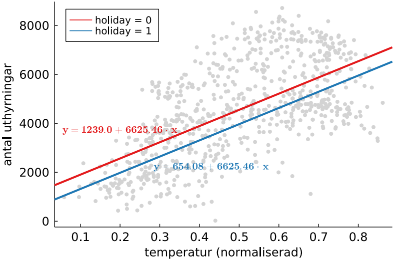
- Varianter av **kodning**: (0,1) eller (−1,1), eller (`true`,`false`).
- Regressionsmodell med binär förklarande variabel:

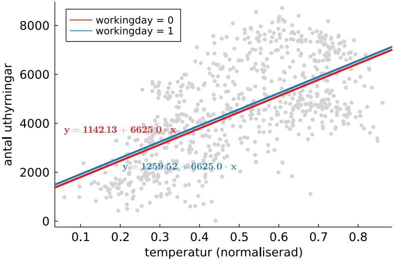
$$y = \alpha + \beta_1 \cdot \text{temp} + \beta_2 \cdot \text{workingday} + \varepsilon$$

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






```
> lmfit = lm(nRides ~ temp + holiday, data = bike)
> regsummary(lmfit, conf_intervals = T, anova = F)

Measures of model fit
------------------------------------------------
  Root MSE         R2     R2-adj
1507.25087    0.39629    0.39464

Parameter estimates
------------------------------------------------
            Estimate Std. Error t value  Pr(>|t|)     2.5 %    97.5 %
(Intercept)  1239.00     161.53  7.6702 5.516e-14    921.87 1556.131
temp         6625.46     304.88 21.7314 4.296e-81   6026.91 7224.007
holiday      -584.92     333.87 -1.7519 8.021e-02  -1240.39   70.552
```

```
> lmfit = lm(nRides ~ temp + workingday, data = bike)
> regsummary(lmfit, conf_intervals = T, anova = F)

Measures of model fit
------------------------------------------------
  Root MSE         R2     R2-adj
1509.43722    0.39454    0.39288

Parameter estimates
------------------------------------------------
            Estimate Std. Error  t value    Pr(>|t|)   2.5 %  97.5 %
(Intercept)  1142.13     177.46  6.43596 2.2248e-10  793.74 1490.53
temp         6625.00     305.62 21.67711 8.7933e-81 6024.99 7225.01
workingday    117.38     120.25  0.97617 3.2931e-01 -118.69  353.46
```

Ännu fler binära förklarande variabler

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

Analysis of variance - ANOVA
------------------------------------------------
       df         SS        MS      F      Pr(>F)
Regr    6 1995735713 332622619 323.77 3.7475e-201
Error 724  743799679   1027348
Total 730 2739535392

Measures of model fit
------------------------------------------------
  Root MSE        R2     R2-adj
1013.58158   0.72849    0.72624

Parameter estimates
------------------------------------------------
             Estimate Std. Error  t value    Pr(>|t|)     2.5 %   97.5 %
(Intercept)  2577.864    252.560 10.20694  6.0523e-23  2082.027  3073.70
temp         6280.856    209.044 30.04567 2.1247e-129  5870.452  6691.26
hum         -2220.634    275.170 -8.07004  2.9221e-15 -2760.861 -1680.41
windspeed   -4363.749    504.412 -8.65115  3.2727e-17 -5354.034 -3373.46
workingday     76.552     83.452  0.91733  3.5928e-01   -87.284   240.39
holiday      -607.036    232.021 -2.61629  9.0743e-03 -1062.551  -151.52
yr           2008.609     75.632 26.55763 4.9489e-109  1860.125  2157.09
```

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:

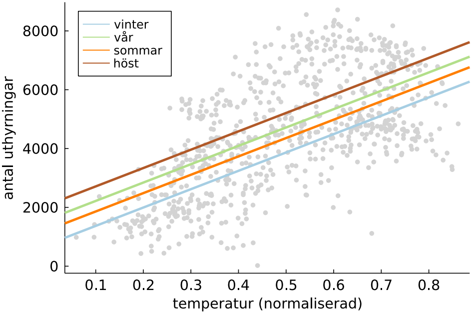
$$y = a + b_1 \cdot \text{temp}$$

- 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




```
Parameter estimates
-----------------------------------------------
              Estimate Std. Error t value   Pr(>|t|)    2.5 %  97.5 %
(Intercept)    745.79     187.48  3.9780 7.6450e-05  377.728 1113.85
temp          6241.35     518.14 12.0456 1.3596e-30 5224.110 7258.58
springTRUE     848.72     197.08  4.3065 1.8870e-05  461.806 1235.64
summerTRUE     490.20     259.01  1.8926 5.8808e-02  -18.294  998.68
fallTRUE      1342.87     164.59  8.1590 1.4872e-15 1019.748 1666.00
```

Cykeluthyrning - säsongsdummies

```
> bike$spring <- bike$season == 2
> bike$summer <- bike$season == 3
> bike$fall <- bike$season == 4
> lmfit = lm(nRides ~ temp + spring + summer + fall, data = bike)
> regsummary(lmfit, conf_intervals = T)

Analysis of variance - ANOVA
------------------------------------------------
       df         SS        MS      F     Pr(>F)
Regr    4 1248576475 312144119 151.99 2.0539e-94
Error 726 1490958917   2053662
Total 730 2739535392

Measures of model fit
------------------------------------------------
  Root MSE         R2     R2-adj
1433.06051    0.45576    0.45276

Parameter estimates
------------------------------------------------
            Estimate Std. Error t value   Pr(>|t|)    2.5 %  97.5 %
(Intercept)   745.79     187.48  3.9780 7.6450e-05  377.728 1113.85
temp         6241.35     518.14 12.0456 1.3596e-30 5224.110 7258.58
springTRUE    848.72     197.08  4.3065 1.8870e-05  461.806 1235.64
summerTRUE    490.20     259.01  1.8926 5.8808e-02  -18.294  998.68
fallTRUE     1342.87     164.59  8.1590 1.4872e-15 1019.748 1666.00
```

Kategoriska variabler med R's faktorvariabler

```
> lmfit = lm(nRides ~ temp + as.factor(season), data = bike)
> regsummary(lmfit, conf_intervals = T)

Analysis of variance - ANOVA
------------------------------------------------
       df         SS        MS      F     Pr(>F)
Regr    4 1248576475 312144119 151.99 2.0539e-94
Error 726 1490958917   2053662
Total 730 2739535392

Measures of model fit
------------------------------------------------
  Root MSE        R2     R2-adj
1433.06051    0.45576    0.45276

Parameter estimates
------------------------------------------------
                  Estimate Std. Error t value   Pr(>|t|)    2.5 %  97.5 %
(Intercept)         745.79     187.48  3.9780 7.6450e-05  377.728 1113.85
temp               6241.35     518.14 12.0456 1.3596e-30 5224.110 7258.58
as.factor(season)2  848.72     197.08  4.3065 1.8870e-05  461.806 1235.64
as.factor(season)3  490.20     259.01  1.8926 5.8808e-02  -18.294  998.68
as.factor(season)4 1342.87     164.59  8.1590 1.4872e-15 1019.748 1666.00
```

F-test för en grupp av förklarande variabler

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

$$H_0 : \beta_{\text{vår}} = \beta_{\text{sommar}} = \beta_{\text{höst}} = 0$$
$$H_1 : \text{någon av } \beta_{\text{vår}}, \beta_{\text{sommar}} \text{ eller } \beta_{\text{höst}} \text{ är skild från noll.}$$

- **Teststatistiska**

$$F = \frac{\left(R^2_{\mathrm{UR}} - R^2_{\mathrm{R}}\right) / r}{\left(1 - R^2_{\mathrm{UR}}\right) / (n - k - 1)}$$

  - R^2_{UR} är R^2 för regressionen **U**tan nollhypotesens **R**estriktioner (de tre säsongsdummies är med i modellen)
  - R^2_{R} 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 $r = 3$.
- $R^2_{\mathrm{UR}} = 0.5354$
- $R^2_{\mathrm{R}} = 0.4609$

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

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

- $F_{\mathrm{obs}} > F_{\mathrm{crit}}$ så nollhypotesen förkastas på signifikansnivån 5%. Det verkar finns en säsonseffekt.