

Föreläsning 9: Multipel linjär regression och modellval

Matias Quiroz¹

¹Statistiska institutionen, Stockholms universitet

VT 2023

- ▶ Multipel linjär regression.
- ▶ Tolkning av multipel linjär regression.
- ▶ Prediktion i multipel linjär regression.
- ▶ Residualanalys.
- ▶ Dummyvariabler i regression.
- ▶ Modellval.
- ▶ Modellval genom korsvalidering.

Fler variabler förklarar mer variation

- ▶ Förra föreläsningen gick vi igenom enkel linjär regression.
- ▶ R-kvadrat är ett mått på **förklaringsgraden**: Hur mycket av variationen i y kan vi fånga med hjälp av x .
- ▶ Exempel: Kroppsfett mot midjemått för 250 män:

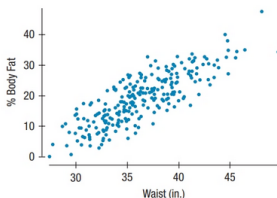


Figure 1: Figur 9.1 i De Veaux et al. (2021).

- ▶ Den räta linjen $\widehat{Body\ Fat} = -42.7 + 1.7Waist$ ger $R^2 = 0.678$, dvs midjemått förklarar nästan 68% av variationen i kroppsfett.
- ▶ Finns det några andra variabler som kan förklara en del av de resterande 32%?

Multipl linjär regression

- ▶ Multipl linjär regression tillåter oss göra en minsta kvadratanpassning när vi har fler än en förklarande variabel.
- ▶ Idén för minsta kvadratanpassningen är samma som förut.
- ▶ Vi vill **miminimera residualkvadratsumman**, men nu har vi fler variabler som predikterar y .
- ▶ Exempel: Om vi har två variabler x_1 och x_2 predikteras y enligt

$$\hat{y} = b_0 + b_1x_1 + b_2x_2. \quad (1)$$

- ▶ Vi vill att minsta kvadratanpassningen minimerar

$$\sum e^2, \quad (2)$$

där $e = y - \hat{y} = y - (b_0 + b_1x_1 + b_2x_2)$.

- ▶ De värden på b_0 , b_1 och b_2 som minimerar (2) används i (1) för att prediktera y .

Multipl linjär regression, forts.

- ▶ Vi kan generalisera idén till godtyckligt många variabler.
- ▶ En **multipl linjär regression** med k variabler, x_1, x_2, \dots, x_k , predikterar y enligt

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k. \quad (3)$$

- ▶ Minsta kvadratanpassningen minimerar

$$\sum e^2, \quad (4)$$

där $e = y - \hat{y} = y - (b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k)$.

- ▶ De värden på $b_0, b_1, b_2, \dots, b_k$ som minimerar (4) används i (3) för att prediktera y .

► Minsta kvadratanpassningens

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

egenskaper:

1. Minimerar residualkvadratsumman i (4).
 2. Residualerna summer till 0, dvs $\sum e = 0$.
 3. Den anpassade linjen går genom punkten $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k, \bar{y})$.
- Formlerna för att räkna ut $b_0, b_1, b_2, \dots, b_k$ är krångliga.
- Vi använder programvara (R!) för räkna ut minsta kvadratanpassningen.

Multipl linjär regression, forts.

- För enkel linjär regression räknar `lm` i R ut b_0 och b_1 .

```
load("Datasets/Bodyfat.Rdata")
lm(Pct.BF ~ Waist, data = Bodyfat)
Call: lm(formula = Pct.BF ~ Waist, data = Bodyfat)
Coefficients: (Intercept) Waist -42.7 1.7
```

- Enkelt att modifiera `lm` för multipl linjär regression. Exempel när vi inkludera längd som en till prediktor.

```
load("Datasets/Bodyfat.Rdata")
lm(Pct.BF ~ Waist + Height, data = Bodyfat)
Call: lm(formula = Pct.BF ~ Waist + Height, data = Bodyfat)
Coefficients: (Intercept) Waist Height -3.101 1.773 -0.602
```

- $b_0 = -3.101$, $b_1 = 1.773$ och $b_2 = -0.602$.

Multipl linjär regression, forts.

- R output av `summary()` för den enkla linjära regressionen

```
> summary(lm(Pct.BF ~ Waist, data = Bodyfat))
```

```
Call:
lm(formula = Pct.BF ~ Waist, data = Bodyfat)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-10.8987  -3.6453   0.1864   3.1775  12.7887
```

```
Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.73413    2.71651  -15.73  <2e-16 ***
Waist         1.69997     0.07431   22.88  <2e-16 ***
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.713 on 248 degrees of freedom
Multiple R-squared:  0.6785,    Adjusted R-squared:  0.6772
F-statistic: 523.3 on 1 and 248 DF,  p-value: < 2.2e-16
```

- Mestadels av outputen förklaras i Del 2 av kursen.
- Vi känner igen de rödmarkerade: **Residualernas fördelningsmått samt min och max, skattningarna b_0 och b_1 , residualernas standardavvikelse s_e , och R-kvadrat R^2 .**

Multipel linjär regression, forts.

- R output av `summary()` för den multipla linjära regressionen

```
> summary(lm(Pct.BF ~ Waist + Height, data = Bodyfat))
```

```
Call:
```

```
lm(formula = Pct.BF ~ Waist + Height, data = Bodyfat)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-11.1692	-3.4133	-0.0977	3.0995	9.9082

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.10088	7.68611	-0.403	0.687
Waist	1.77309	0.07158	24.770	< 2e-16 ***
Height	-0.60154	0.10994	-5.472	1.09e-07 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.46 on 247 degrees of freedom
```

```
Multiple R-squared: 0.7132, Adjusted R-squared: 0.7109
```

```
F-statistic: 307.1 on 2 and 247 DF, p-value: < 2.2e-16
```

- Mestadels av outputen förklaras i Del 2 av kursen.
- Vi känner igen de rödmarkerade: **Residualernas fördelningsmått samt min och max**, **skattningarna** b_0 , b_1 och b_2 , **residualernas standardavvikelse** s_e , och **R-kvadrat** R^2 .

Multipl linjär regression, forts.

- ▶ Notera att den multipla linjära regressionen förklarar mer av variationen:
 1. Större R-kvadrat: $R^2 = 0.6785$ (enkel) jämfört med $R^2 = 0.7135$ (multipl).
 2. Mindre residualstandardavvikelse: $s_e = 4.713$ jämfört med $s_e = 4.46$.
- ▶ Man kan visa att R^2 **alltid blir större i en multipl regression med fler variabler**.
- ▶ Man kan också visa att s_e **alltid blir mindre i en multipl regression med fler variabler**.
- ▶ Om vi inkluderar en tredje förklarande variabel i regressionen ovan, hade dess R^2 varit ännu större och s_e varit ännu mindre.
- ▶ Intuitivt kan man förstå fenomenet som att det sämsta som kan hända är att den nya förklarande variabeln som inkluderas inte förklarar någon variation.
- ▶ Det ovannämnda skulle inträffa om den nya förklarande variabeln var okorrelerad med responsen y .

Tolkning av multipel linjär regression

- Minsta kvadratanpassningen för vår regression med två förklarande variabler

$$\widehat{Body\ Fat} = -3.101 + 1.773Waist - 0.602Height.$$

- Ekvationen tycks säga att det finns ett negativt samband mellan kroppsfett och längd.
- Låt oss kolla på punktdiagrammet mellan kroppsfett och längd:

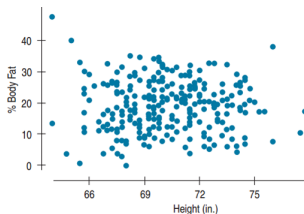


Figure 2: Figur 9.2 i De Veaux et al. (2021).

- Punktdiagrammet verkar inte visa något samband. Vad händer?

Tolkning av multipel linjär regression, forts.

- Den anpassade modellen: $\widehat{Body\ Fat} = -3.101 + 1.773Waist - 0.602Height$.
- Anpassningen av b_2 tar också hänsyn till vad midjemåttet är.
- Antag att vi håller midjemåttet x_1 konstant vid 37 (runt \bar{x}_1). Förväntar vi oss ett negativt samband?
- Ja, längre män bör i genomsnitt ha lägre kropps fett jämfört med kortare män givet att de har samma midjemått.
- Negativt samband för män som har midjemått 36-38:

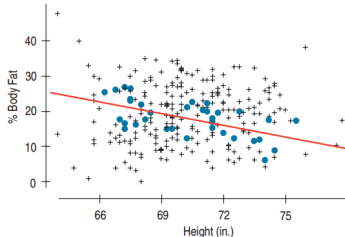


Figure 3: Figur 9.3 i De Veaux et al. (2021).

Tolkning av multipel linjär regression, forts.

- ▶ I en multipel linjär regression måste vi alltid tolka sambanden (positivt/negativt) **givet värden på de andra variablerna**.
- ▶ Detta är en **betingad tolkning**.
- ▶ När vi försökte tolka b_2 från figuren nedan så blev det fel, eftersom figuren visar ett marginellt (obefintligt) samband mellan kropps fett och längd:

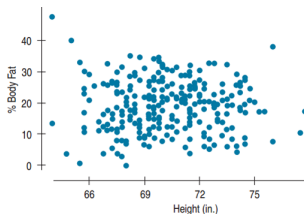


Figure 4: Figur 9.2 i De Veaux et al. (2021).

- ▶ Marginellt här betyder att vi betraktar sambandet mellan kropps fett och längd oberoende av midjemått.

Tolkning av multipel linjär regression, forts.

- ▶ Låt oss betraka ett annat exempel då vi måste vara försiktiga med hur vi tolkar en multipel linjär regression.
- ▶ En regression predikterar huspriser med hjälp av antal sovrum enligt

$$\widehat{Price} = 338975 + 40234 \text{Bedroom},$$

och visar ett rimligt positivt samband mellan variablerna.

- ▶ Hus med ett extra sovrum tenderar att i genomsnitt säljas för \$40234 mer.
- ▶ Tolkningen av en enkel linjär regression är aldrig svår.

- ▶ När vi också inkluderar boytan som prediktor blir prediktionen

$$\widehat{Price} = 308100 + 135Living\ Area - 43347Bedroom. \quad (5)$$

- ▶ Koefficienten för *Bedroom* i (5) visar inte ett negativt samband mellan pris och antal sovrum, dvs att hus med fler sovrum säljs till lägre pris.
- ▶ Tolkningen måste alltid **hålla dom andra variablerna konstanta**.
- ▶ Om boytan är konstant, kan det lägre priset i lägenheter med fler sovrum förklaras av något av följande:
 - ▶ Andra boytor i hemmet (vardagsrum, kök, osv) måste vara mindre.
 - ▶ Sovrummen är mindre.
- ▶ Tänk alltid efter när man tolkar multipel linjär regression, speciellt om sambanden ändras kraftigt efter att ha inkluderat en variabel.

Prediktion i multipel linjär regression

- ▶ Prediktion är lika enkelt som för enkel linjär regression.
- ▶ Vi vill sälja en lägenhet på 904 square feet (ca 82 kvadratmeter) och som har 3 sovrum.
- ▶ Enligt vår modell kan vi förvänta oss att en sådan lägenhet i genomsnitt säljs

$$\hat{y} = 308100 + 135 \cdot 904 - 43347 \cdot 3 = 300099,$$

dvs ca \$300000.

- ▶ Som vanligt gäller att inte prediktera för variabelvärden x som är utanför intervallet för de x -värden som har använts för att anpassa modellen.

Förutsättningar och antaganden i multipel linjär regression

- ▶ Både y och alla x_1, x_2, \dots, x_k måste vara numeriska variabler.
- ▶ y måste förhålla sig (approximativt) linjärt till vardera prediktor.
- ▶ Inga uppenbara outliers — kan påverka minsta kvadratanpassningen. Anpassa linjen utan outliers för att kontrollera att resultaten blir ungefär desamma.
- ▶ Spridningen för y beror inte på x_1, x_2, \dots, x_k . Residualernas varians måste vara konstant.
- ▶ Residualernas bör vara (approximativt) normalfördelade.
- ▶ Dessa kan valideras genom en residualanalys.

- Den anpassade multipla linjära regressionen är

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k.$$

- En förutsättning för att använda multipel linjär regression är att den linjära modellen måste vara trovärdig, dvs anpassa observerade data.
- “All models are wrong, but some are useful” – George E. P. Box.
- Kom ihåg: Om modellen beskriver data på ett adekvat sätt, så kommer residualerna inte ha något tydligt mönster i sig. De beter sig slumpmässigt.
- I Föreläsning 8 lärde vi oss att en förutsättning för att residualerna inte ska visa ett uppenbart mönster är att y och x måste förhålla sig linjärt.
- I multipel linjär regression kan vi göra parvisa punktdiagram mellan y och vardera förklarande variabel för att kolla detta antagande.
- Om de inte förhåller sig linjärt så kan vi transformera variablerna för att få ett linjärt förhållande. **Ladder of powers** från Föreläsning 8.

- ▶ Residualernas standardavvikelse kan räknas enligt

$$s_e = \sqrt{\frac{\sum e^2}{n - k - 1}}.$$

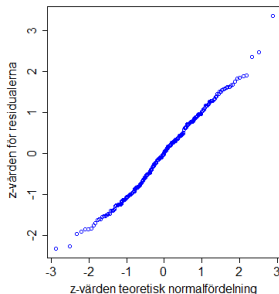
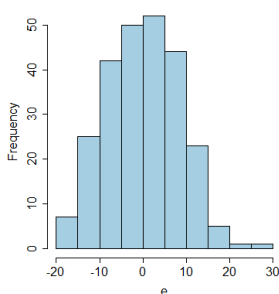
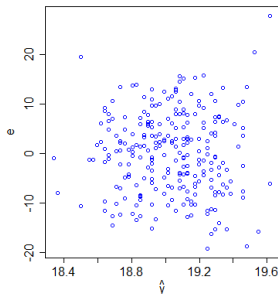
- ▶ Varför delar vi med $n - k - 1$ istället för n , $n - 1$, $n - 2$? Vi behöver lära oss mer statistik innan vi kan förklara det.
- ▶ För att räkna samplingfördelningar för b_1, \dots, b_k (Del 2 av kursen) behöver vi fler modellantaganden som medför:
 1. **Residualerna är normalfördelade.**
 2. **Residualernas varians är konstant**, dvs beror inte på x .
- ▶ Vi kan undersöka 1. genom 68–95–99.7 regeln eller en normalfördelningsplot.
- ▶ Vi kan undersöka 2. genom att plotta e mot \hat{y} och se om spridningen är konstant.
- ▶ I enkel linjär regression plottade vi e mot x . I multipel linjär regression kan vi ha många x , därför enklare att plotta e mot \hat{y} istället.

Residualanalys, forts.

- Låt oss göra en residualanalys för modellen

$$\widehat{Body\ Fat} = -3.101 + 1.773Waist - 0.602Height.$$

- Residualanalysen visar att modellen är trovärdig.



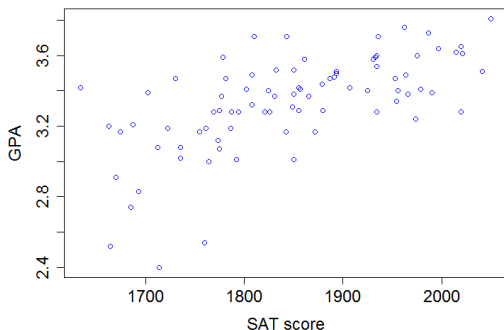
- Kommentarer:

- Residualernas spridning är någorlunda konstant.
- Histogrammet och normalfördelningsploten visar att data är approximativt normalfördelade.

- ▶ I många problem är det svårt att få antagandena att stämma.
- ▶ Uppgift 5.4 i Inlämningsuppgift 1 är ett exempel.
- ▶ “All models are wrong, but some are useful” – George E. P. Box.
- ▶ Så länge en regressionsmodell kan ge en bättre prediktion än prediktionen \bar{y} så är den “useful” i någon mening.
- ▶ Viktigt att kommunicera modellens brister.

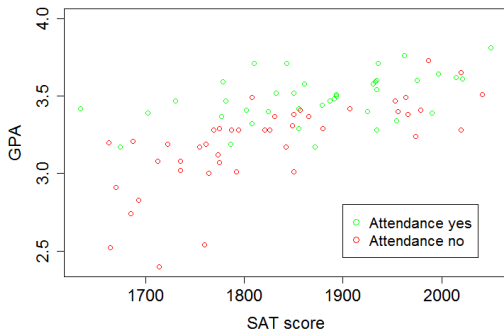
Dummyvariabler i regression

- ▶ En av förutsättningarna för regression är att y och alla x_1, x_2, \dots, x_k måste vara numeriska variabler.
- ▶ Ibland vill man ha en **kategorisk variabel som prediktor**.
- ▶ Låt oss prediktera snittbetyg (GPA) i USA med hjälp av 84 universitetsstudenters SAT score (amerikanska varianten av högskoleprovet).
- ▶ Båda dessa variabler är numeriska. Ett positivt samband verkar föreligga:



Dummyvariabler i regression, forts.

- ▶ Antag nu att vi också har en variabel `attendance`, som har utfallet `yes` om studenten deltagit i minst 75% av föreläsningar och `no` annars.
- ▶ Samma som föregående figur, men nu visar färgen vad studenten hade för `attendance`:



- ▶ Sambandet mellan GPA och SAT score verkar skilja sig beroende på `attendance`. Hur kan vi inkludera denna information i regressionen?

Dummyvariabler i regression, forts.

- ▶ Vi kan skapa en såkallad dummyvariabel (**dummy/indicator variable** på engelska).
- ▶ En dummyvariabel kodar en kategorisk variabel med två utfall som en numerisk variabel med utfallet 0 eller 1.
- ▶ Den omkodade variabeln används i en multipel linjär regression tillsammans med de andra variablerna (SAT score i vårt fall).
- ▶ Kodningen av attendance="yes" som 1 och attendance="no" som 0, eller tvärtom, är godtycklig men påverkar tolkningen.
- ▶ Vi väljer attendance="yes" som 1 och attendance="no" som 0.

```
load("Datasets/GPA.Rdata")  Attendance coded as yes = 1 and no = 0
lm(GPA ~ SAT + attendance, data = GPA)

Call: lm(formula = GPA ~ SAT + attendance, data = GPA)
Coefficients: (Intercept) SAT attendance 0.6439 0.0014 0.2226
```


Dummyvariabler i regression, forts.

- R output av `summary()` för den enkla linjära regressionen

```
> summary(lm(GPA ~ SAT + attendance, data = GPA))
```

```
Call:
```

```
lm(formula = GPA ~ SAT + attendance, data = GPA)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-0.64311	-0.06820	0.01251	0.11787	0.31531

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.643850	0.358205	1.797	0.076 .
SAT	0.001400	0.000196	7.141	3.60e-10 ***
attendance	0.222644	0.040842	5.451	5.27e-07 ***

```
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.1813 on 81 degrees of freedom  
Multiple R-squared:  0.5654,    Adjusted R-squared:  0.5547  
F-statistic: 52.7 on 2 and 81 DF,  p-value: 2.193e-15
```

- Ungefär 56% av variationen i GPA förklaras av SAT och attendance.
- Hur tolkar vi coefficienten för attendance, dvs $b_2 = 0.2227644 \approx 0.223$?

Dummyvariabler i regression, forts.

► Prediktionen: $\widehat{GPA} = 0.64385 + 0.0014SAT + 0.223attendance$.

► Prediktionen för student 1 som har $attendance=0$ ("No")

$$\widehat{GPA}^{(1)} = 0.64385 + 0.0014SAT + 0.223 \cdot 0 = 0.64385 + 0.0014SAT.$$

► Prediktionen för student 2 som har $attendance=1$ ("Yes")

$$\widehat{GPA}^{(2)} = 0.64385 + 0.0014SAT + 0.223 \cdot 1 = 0.64385 + 0.0014SAT + 0.023.$$

► Givet att **båda studenterna har samma SAT**,

$$\widehat{GPA}^{(2)} = \underbrace{0.64385 + 0.0014SAT}_{\widehat{GPA}^{(1)}} + 0.223 \cdot 1 = \widehat{GPA}^{(1)} + 0.023.$$

► Givet samma SAT, tenderar en student som har deltagit i minst 75% av föreläsningarna ($dummy=1$) att i **genomsnitt** ha 0.023 högre GPA jämfört med en student som deltagit mindre än 75% ($dummy=0$).

Dummyvariabler i regression, forts.

- ▶ Tolkningen är utifrån kodningen av dummyvariabeln (dvs vilken kategori som är 0 eller 1)!
- ▶ Viktigt att hålla koll på hur man har kodat dummyvariabeln!
- ▶ Allmän tolkning av koefficienten b för en dummyvariabel är följande.
- ▶ **Givet alla andra variablerna lika**, kommer prediktionen för en observation vars dummy=1 vara b större ($b > 0$), eller b mindre ($b < 0$), än prediktionen för en observation vars dummy=0.

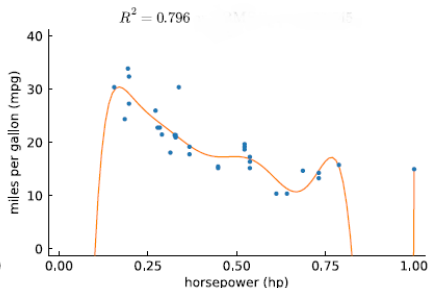
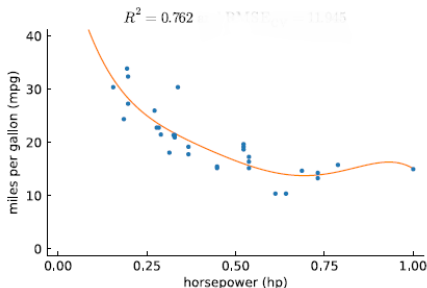
- Den multipla linjära regressionsmodellen predikterar enligt

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k.$$

- Är alla förklarande variabler x_1, x_2, \dots, x_k viktiga för att prediktera y ?
- Resonemang: Bättre att inkludera fler än för få, använd alla variabler!
- Det finns ett problem med resonemanget: Överanpassning av data (**overfitting the data** på engelska).
- Enkelt förklarat: Ju mer "flexibel" modellen är, desto mer av variationen i y kan fångas.
- Låter som en fantastisk egenskap vid en första anblick. Vi vill ju fånga variationen i y ! Vad är problemet?

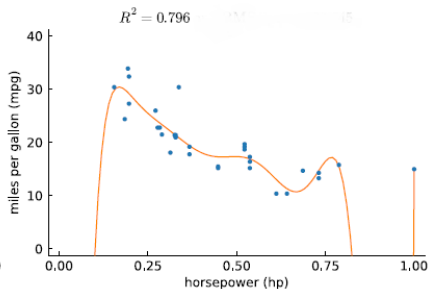
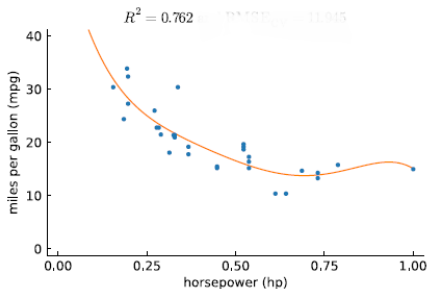
Modellval, forts.

- ▶ Enklast att förklara överanpassning med polynom regression (en icke-linjär regression som täcks i SDA II). Teaser: Villanis widget.
- ▶ Miles per gallon (mpg) mot horsepower (hp) för olika bilar:



- ▶ Modellen till höger är mer flexibel än den till vänster.
- ▶ Den flexibla modellen har högre R^2 , dvs fångar mer av variationen i y .
- ▶ Vilken modell skulle ni ha valt? Varför?

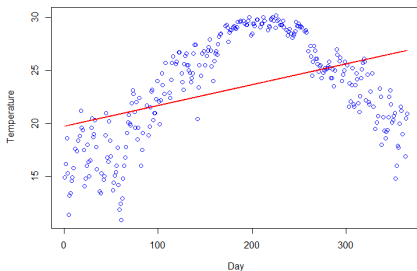
Modellval, forts.



- ▶ Antag att vi vill prediktera y för $x = 0.85$ (horsepower skalad till intervallet $[0, 1]$). Den mer flexibla modellen ger en galen prediktion!
- ▶ En modell sägs ha en generaliseringsförmåga (**generalization** på engelska) om modellen har en förmåga att på ett tillförligt sätt prediktera y med hjälp av x -värden som **inte användes när modellen anpassades**.
- ▶ När en alltför flexibel modell har en dålig generaliseringsförmåga säger vi att **modellen överanpassar data**.

Modellval, forts.

- ▶ En modell kan också underanpassa data (**underfitting the data** på engelska) och därmed ha en dålig generaliseringsförmåga.
- ▶ Exempel: 300 dagars temperaturer i en japansk stad under ett år (65 saknas):



- ▶ Antag att vi vill prediktera y för $x = 229$ (saknas i data). Galen prediktion!
- ▶ Till skillnad från överanpassning, ger underanpassning också dåliga prediktioner för x -värden som användes när modellen anpassades.
- ▶ När en alltför enkel modell ger en dålig anpassning av observationerna säger vi att **modellen underanpassar data**.

Modellval i linjär regression, forts.

- ▶ Finns många tänkbara kriterium på vad som anses vara en bra modell.
- ▶ Ett sådant är att den bör ha god generaliseringsförmåga, dvs modellen varken underanpassar eller överanpassar data.
- ▶ Att studera olika modeller generaliseringsförmåga är en stor grej inom machine learning¹.
- ▶ Att modellera icke-linjärt är ett sätt att få en mer flexibel modell.
- ▶ I en multipel linjär regression är ett annat sätt att få en mer flexibel modell att inkludera fler x variabler.
- ▶ Antag att vi har två multipla regressionsmodeller,

$$\hat{y}^{(1)} = b_0^{(1)} + b_1^{(1)} x_1 + b_2^{(1)} x_2$$

$$\hat{y}^{(2)} = b_0^{(2)} + b_1^{(2)} x_1 + b_2^{(2)} x_2 + b_3^{(2)} x_3 + b_4^{(2)} x_4.$$

- ▶ Kan vi använda R^2 för att avgöra vilken av dom som är bäst?

¹Påminner om kursen Maskininlärning på masternivå.

Modellval i linjär regression, forts.

- ▶ Nej, R^2 blir alltid större när vi inkluderar fler variabler i en regression!
- ▶ R^2 tar inte hänsyn till överanpassning.
- ▶ För att motverka detta kan man använda justerat R-kvadrat (**adjusted R^2** på engelska)

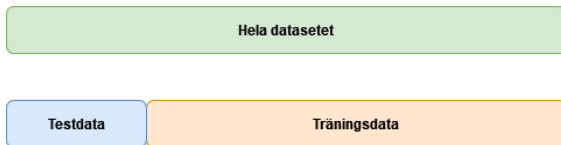
$$R_{\text{adj}}^2 = 1 - (1 - R^2) \frac{n-1}{n-k-1} \left(= 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)} \right),$$

där k är antalet förklarande variabler och n är antal observationer.

- ▶ $\bar{R}^2 < R^2$. Kan bli negativ om R^2 är liten.
- ▶ Ett annat verktyg för modellval är korsvalidering (**cross validation** på engelska).
- ▶ Korsvalidering är mycket populärt i maskininlärning.

Modellval genom korsvalidering, forts.

- ▶ Underliggande idé: Utvärderar modellens generaliseringsförmåga genom att prediktera observationer som inte användes när modellen skattades.
- ▶ Data delas upp i två delmängder:
 - ▶ Träningsdata (**training data** på engelska): Data som används för att anpassa modellen.
 - ▶ Testdata (**test data** på engelska): Data som används för att utvärdera modellens prediktionsförmåga.
- ▶ Exempel med 75% som träningsdata och 25% som testdata.



- ▶ Om data ligger i ordning bör observationerna sorteras slumpmässigt innan uppdelning så båda mängderna är representativa för stickprovet.

Modellval genom korsvalidering, forts.

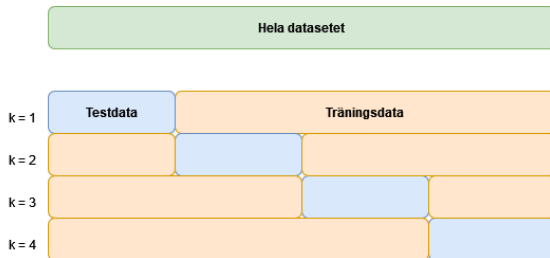
- ▶ Ett mått på prediktionsförmågan är att räkna residualkvadratsumman för de n_{test} testobservationerna,

$$\text{SSE}_{\text{test}} = \sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2.$$

- ▶ Notera
 - ▶ Vi summerar enbart över testobservationerna.
 - ▶ \hat{y}_i är en prediktion baserat på en anpassning **enbart på träningsdata**.
 - ▶ Om anpassningen har överanpassat träningsdata så kommer SSE_{test} vara stor.
- ▶ Att dela upp data enligt ovan och välja modellen med lägst SSE_{test} har en nackdel...
- ▶ ... Vi skulle helst vilja att varje observation får möjlighet att ingå i både träningsdata och i valideringsdata (dock inte samtidigt).
- ▶ Korsvalidering!

Modellval genom korsvalidering, forts.

- Korsvalideringsuppdelning med $K = 4$ så kallade **folds**:



- Notera att vi fick $K = 4$ folds pga uppdelningen 75% och 25%. Vid till exempel 90% och 10% hade vi fått $K = 10$ folds.
- Korsvalideringen SSE för alla observationer.

$$SSE_{cv} = SSE_{test}^{(1)} + SSE_{test}^{(2)} + SSE_{test}^{(3)} + SSE_{test}^{(4)}.$$

- **Mean squared error** (MSE) för korsvalideringen $MSE_{cv} = \frac{SSE_{cv}}{n}$.

Modellval genom korsvalidering, forts.

- ▶ Vanligt att rapportera korsvalideringens **root mean square error** (RMSE)

$$\text{RMSE}_{\text{cv}} = \sqrt{\frac{\text{SSE}_{\text{cv}}}{n}}.$$

- ▶ Antag att vi vill jämföra M stycken modeller för ett givet dataset.
- ▶ Utför korsvalidering på varje modell separat. Notera att man **måste använda samma korsvalideringsuppdelning av data** för alla modellerna.
- ▶ Proceduren ger en RMSE_{cv} för varje modell, dvs

$$\text{RMSE}_{\text{cv}}^{(1)}, \text{RMSE}_{\text{cv}}^{(2)}, \dots, \text{RMSE}_{\text{cv}}^{(M)}.$$

- ▶ Välj den modell som har lägst RMSE_{cv} .
- ▶ Ni kommer att få praktisk erfarenhet av korsvalidering i Lab 4 och Inlämningsuppgift 1.

De Veaux, R. D., Velleman, P., and Bock, D. (2021). *Stats: Data and Models*. Pearson, Harlow, United Kingdom, fifth edition.