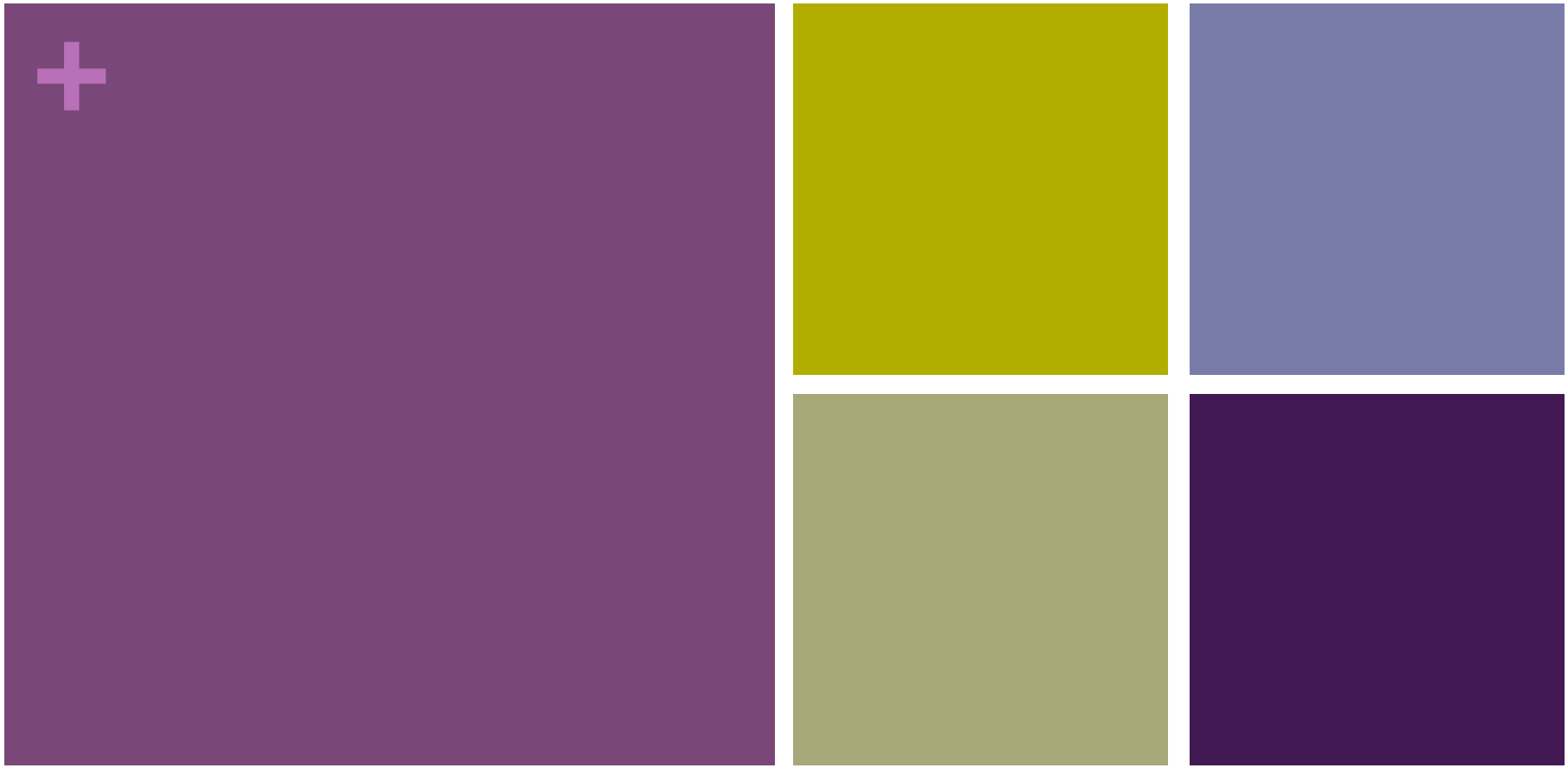


Linjär regressionsanalys

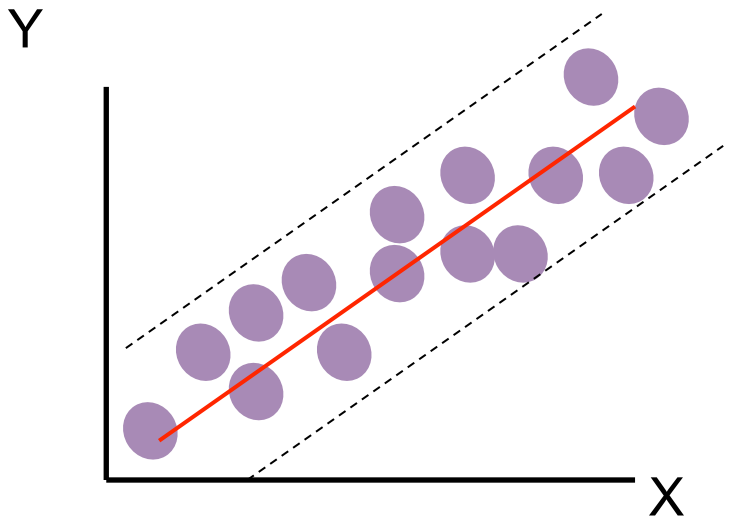
Wieland Wermke

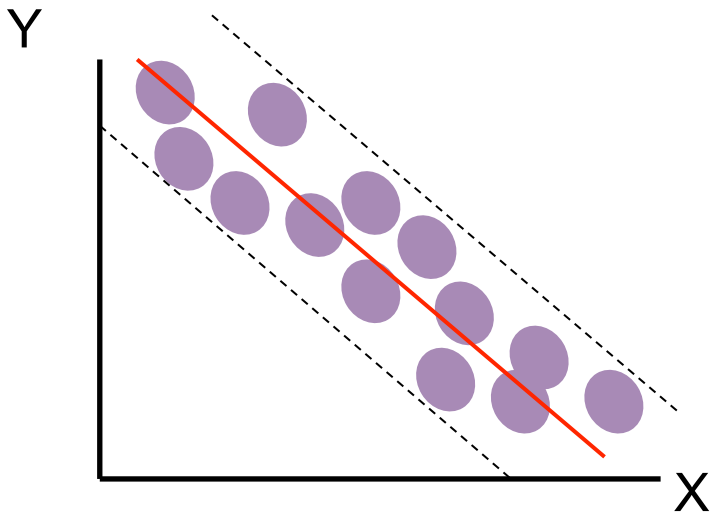
Regressionsanalys

- Analys av samband mellan variabler (x,y)
- Ökad kunskap om x (oberoende variabel) leder till ökad kunskap om y (beroende variabel)
- Utifrån sina data försöker man hitta en ”förutsägelse ekvation”, som kan ge oss bäst möjliga gissning
  - **Detta gör datorn för oss!**
- Enkel linjär regression liknar korrelation
- Obs! Ingen självklar kausalitet i sig, men ett riktat samband!

Enkel linjär regression

- Hitta en linje (i ett koordinatsystem) som beskriver sambandet mellan x (ov) och y (bv) på bästmöjliga sätt

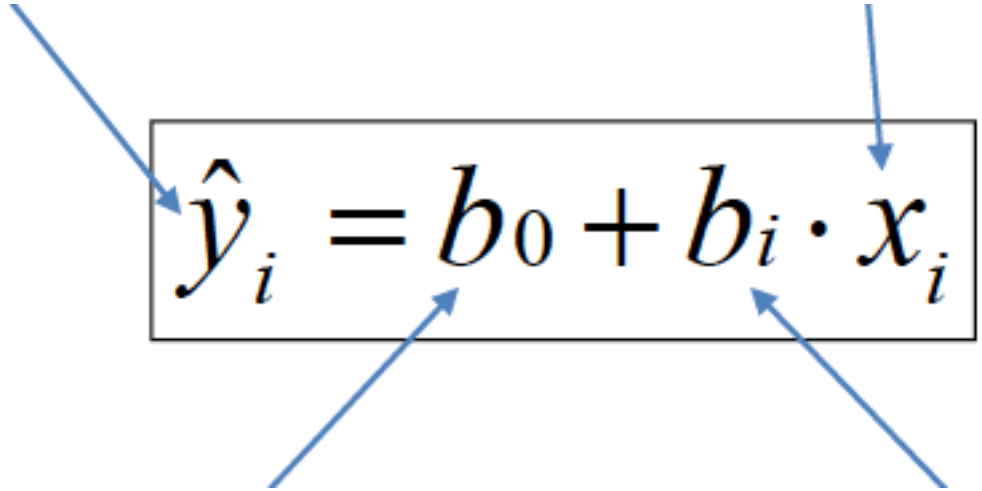




Linjär regression - Formel

Gissade värd utifrån vårt linje

x värde (ov) av person i



Intercept, konstant, här möter linjen y-axeln, dvs. värdet, när x=0

Lutning, genomsnittlig förändring av y, när x förändras med 1 enhet

Standardiserat vs. inte standardiserat?

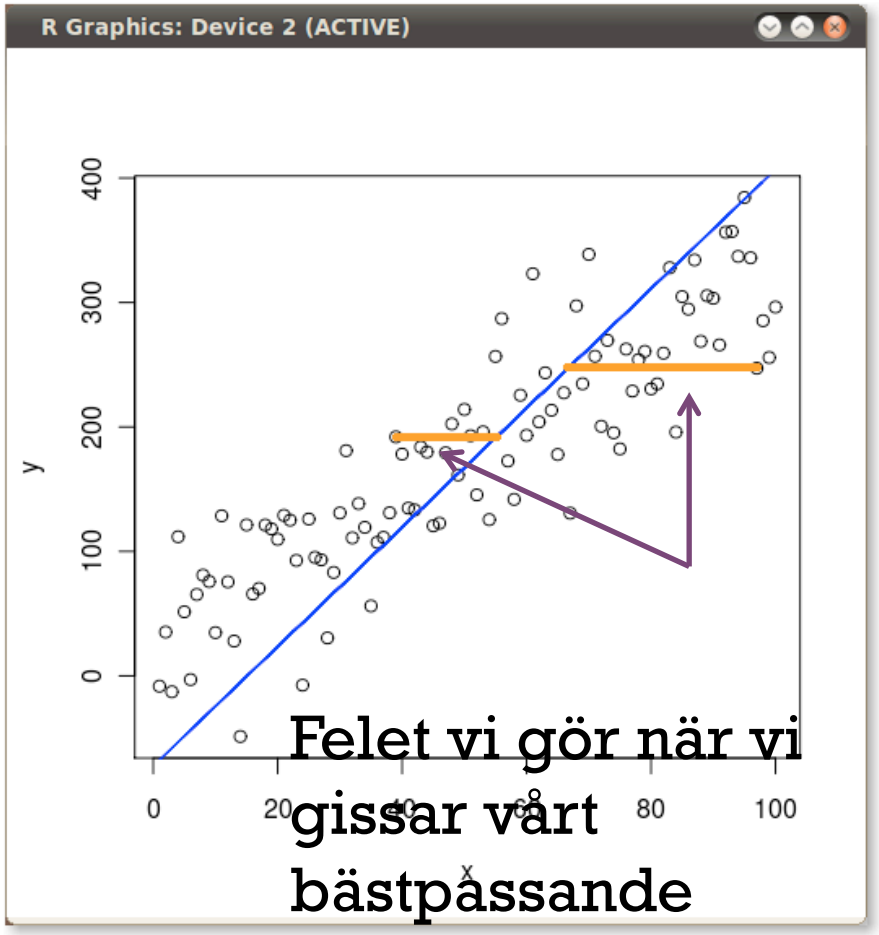
- Lutningen (b) = förändring av beroende variabel, om den beroende variabel förändrar sig med 1
- Det betyder att b beror på vilken dimension man är ut efter (cm, kilo, betyg, poäng)
- Även här kan man standardisera för att jämföra olika lutningar med olika dimensioner (medel= 0, sd=1)

$$\beta_j = b_j \cdot \frac{s_{x_j}}{s_y}$$

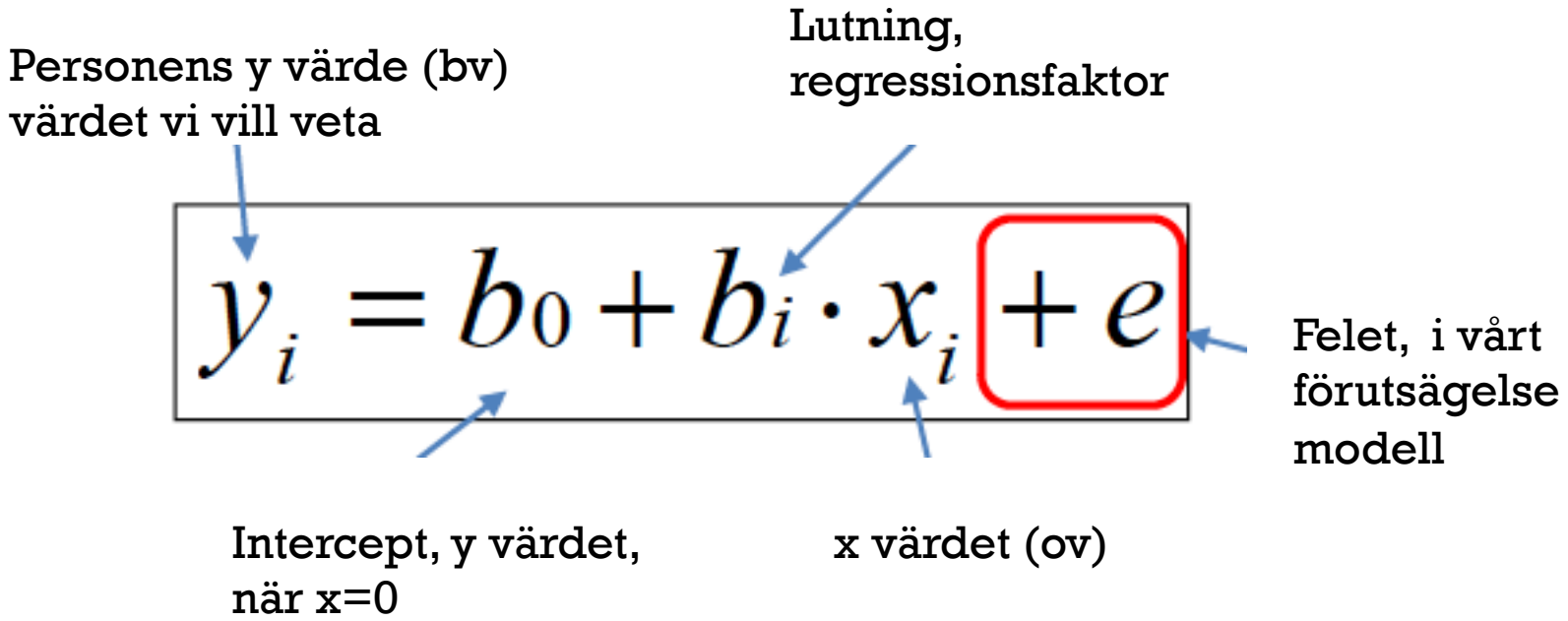
- Beta (β) liknar korrelationskoefficienten (t.ex. Pearson's r) och kan ha värde mellan -1 och 1

Minstkvadratmetoden

- **Så beräknar man formeln:**
- I våra data (punktmolnen) försöker vi hitta linjen som är närmast alla punkter samtidigt (så bra det går)
- För att att hantera ”–/+” (se variansen), kvadrera man varje värde
  - Stora värde tas större hänsyn till än små (modellen blir mer precis med små värde)

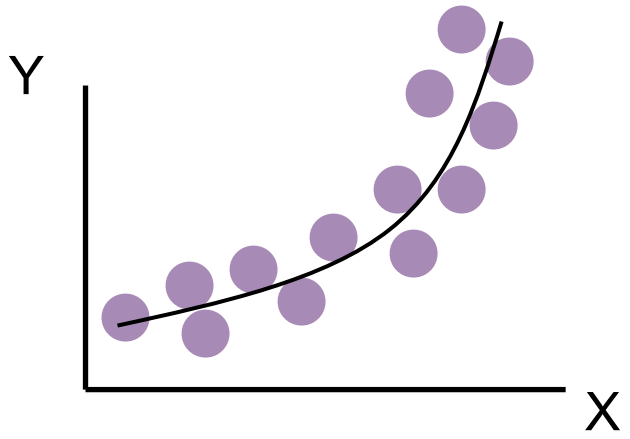


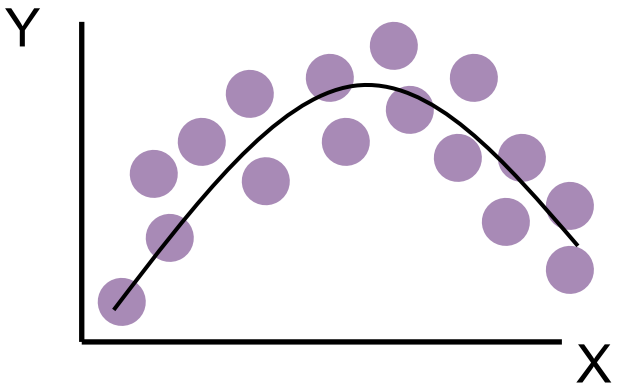
Linjär regression: Formel



Andra samband?

- Linjära samband är inte de enda, det finns andra: Linjär regression fungerar bara om
  - våra data är hyfsat normalfördelade (kan vara ett problem i mycket små urval)
  - Om det finns en linjär samband

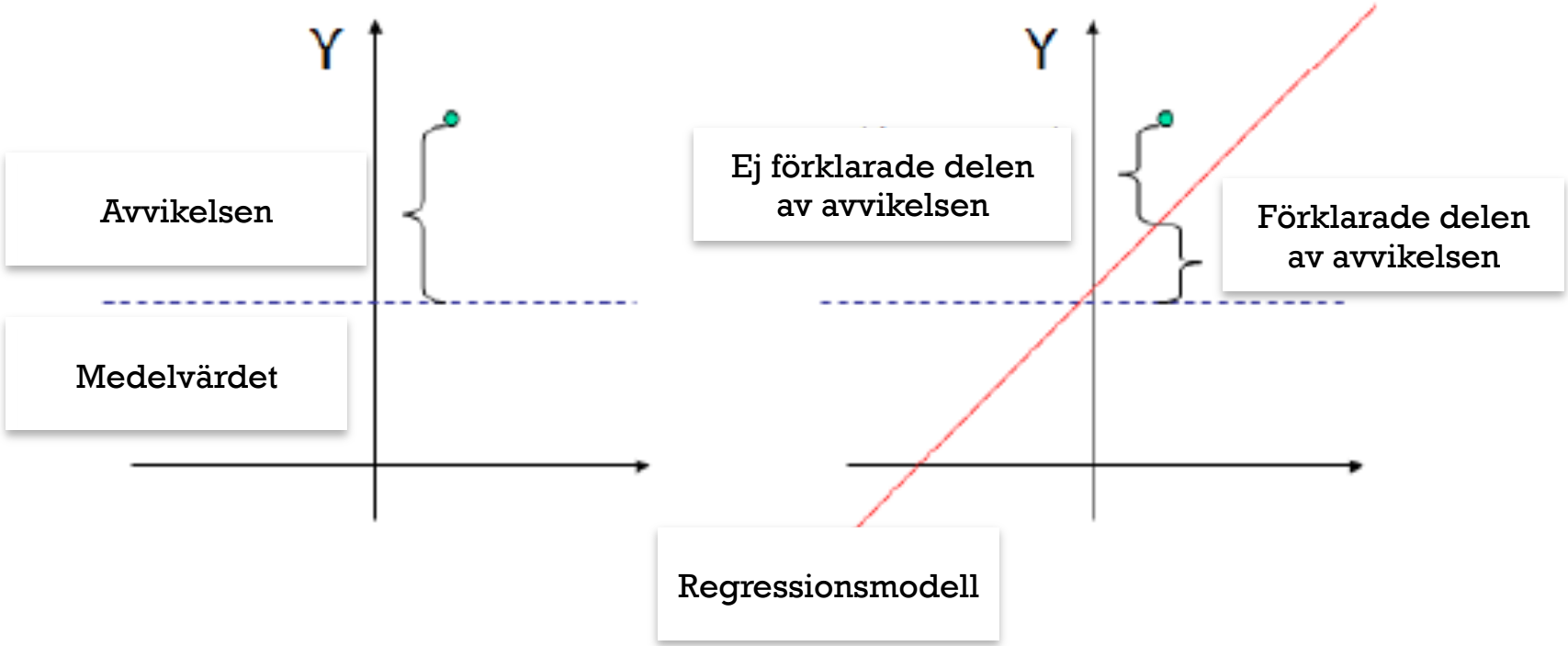




Hur väl fungerar vår modell? R^2

- …säger hur mycket av variansen (i procent) av vår beroende variabel y (avvikelse från medelvärdet) kan förklaras genom våra oberoende variabler
- → säger, hur väl vår modell (vår förutsägelse, gissning) fungerar

Variansförklaring



Att justera R^2

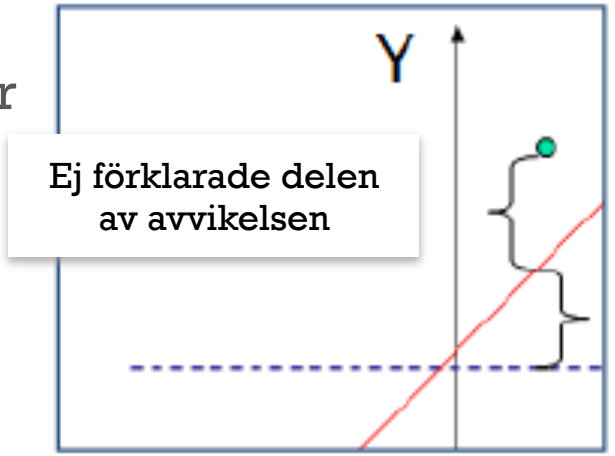
- Genom att ta in hur många variabler som helst in i modellen kan man
- Ju mer variabler, ju större R^2
- Man justera matematiskt i relation till storleken av urvalet
- Ju större N ju mindre faran att en ojusterat R^2 anger för stora värde
- SPSS/PSPP kan stegvis ta in flera variabler och räknar ut hur R^2 förändrar sig

Signifikans av modellen

- **T test** provar var en variabels oberoende statistiskt signifikans
  - Nollhypotesen = Den enstaka oberoende variabeln har inte någon påverkan på beroende variabeln utanför detta urvalet/ Påverkan är slumpmässig
  - Alternativhypotesen = Den enstaka oberoende variabeln har någon påverkan på beroende variabeln utanför detta urvalet
- **F test** provar hela modellens signifikans
  - Nollhypotesen = alla i modellen ingående oberoende variabler tillsammans (dvs. hela modellen) har har inte något påverkan på beroende variabeln/påverkan är slumpmässig
  - Alternativhypotesen= alla i modellen ingående oberoende variabler tillsammans (dvs. hela modellen) har har inte något påverkan på beroende variabeln

Standardfel av modellen

- Detta är standardavvikelsen av vår fel, dvs. detta som inte förklaras av modellen (typ motsatsen av $R^{2)}$
- Beskriver spridningen av våra y värde (BV) kring vår regressionslinje
- Anger hur mycket de ”sanna” värde avviker från vårt modell
- Ju mindre värdet, ju bättre vårt förutsägelse
- Bra för att jämföra olika modeller med varandra



Exempel

- Frågan: Minskar motivationen för statistik kursen ointresset för ämnet statistik?

Model Summary (Quantitative methods and results are rather boring)

R	*R Square*	*Adjusted R Square*	*Std. Error of the Estimate*
.33	.11	.07	.77

ANOVA (Quantitative methods and results are rather boring)

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Regression	1.99	1	1.99	3.33	.079
Residual	16.71	28	.60		
Total	18.70	29			

Coefficients (Quantitative methods and results are rather boring)

	Unstandardized Coefficients		*Standardized Coefficients*		
	B	*Std. Error*	*Beta*	*t*	*Sig.*
(Constant)	3.27	.66	.00	4.99	.000
Workshop motivation (0-10)	-.15	.08	-.33	-1.83	.079

Variansförklaring av modellen (7%)

Signifikansvärde av hela modellen

Genomsnittligt fel vi gör med vårt modell (måttligt)

Minskning av ointresse i ämnet (1-4), om motivationen ökas med 1 =-0,15

Värdet av ointresse när motivation är 0

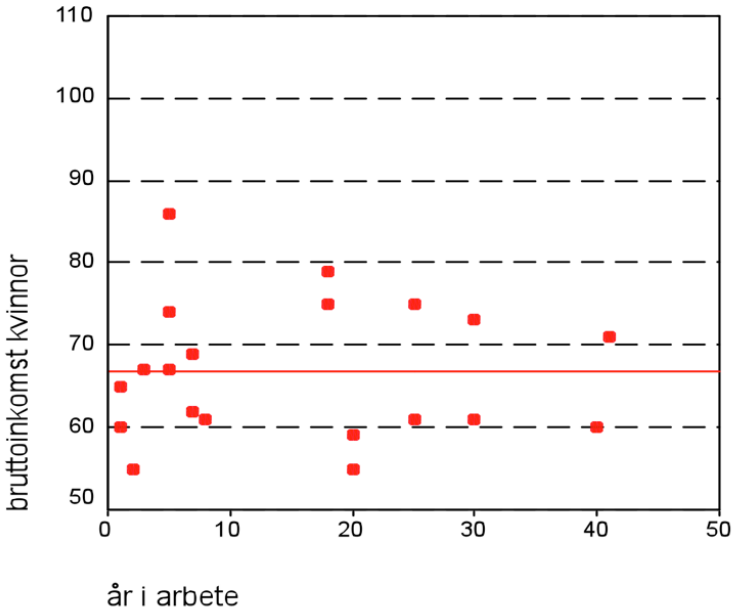
Standardiserat regressionskoefficient (tolkas lika som Pearsons r)=svagt negativt samband)

Signifikansvärde av OV (motivation)

Multivariata linjära samband

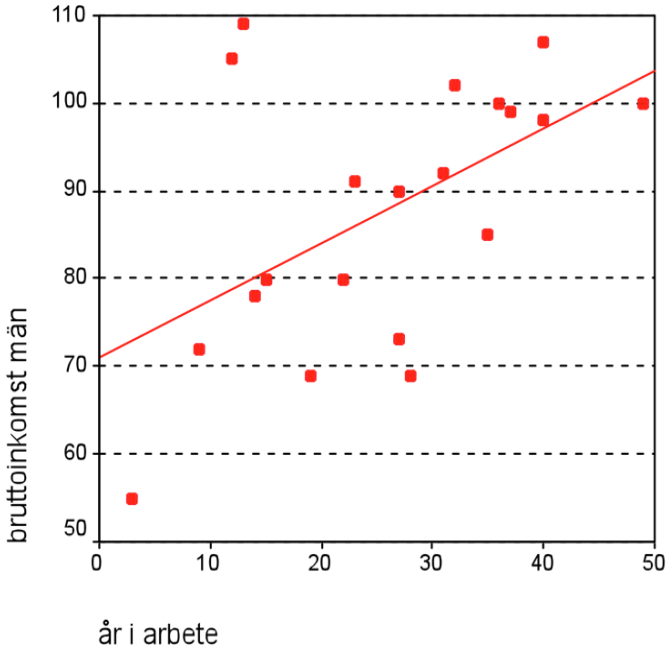
Problem med bivariata samband och andra variabler

- Vad händer i ett observerat samband om man tar hänsyn till en **tredje variabel**?
  - **t.ex. samband r (år yrkeserfarenheter → lön): 0.3**
- Ursprungssambandet kan försvinna eller modereras, om man beaktar en tredje variabel, som t.ex. kön
- exempel samband för kvinnor r=0.00!



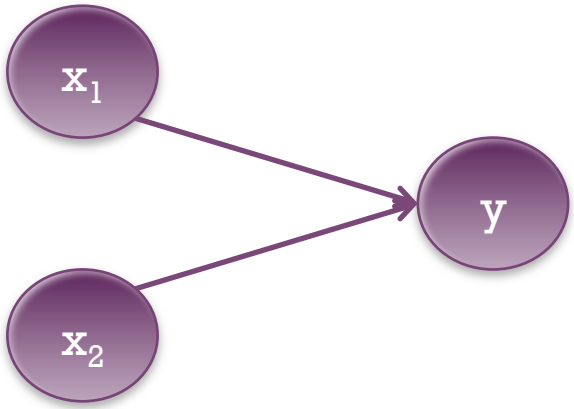
Problemet med bivariata samband och en tredje variabel

- Exempel samband för män r=0.52
- Strategi: **konstanthållning**
  - fundera över vilka variabler som också kan påverka
  - dela upp grupperna efter den variabeln och räkna för båda gruppern separat



Enkel och multipel linjär regression

Enkel linjär regression



Multipel linjär regression

Multivariata regressioner

- MR beräknar olika koefficienter när de andra inflyttade koefficienterna har hållits konstanta (över alla observationsenheter)

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_i x_i$$

Förväntade värde enl. Minstkvadrat modell

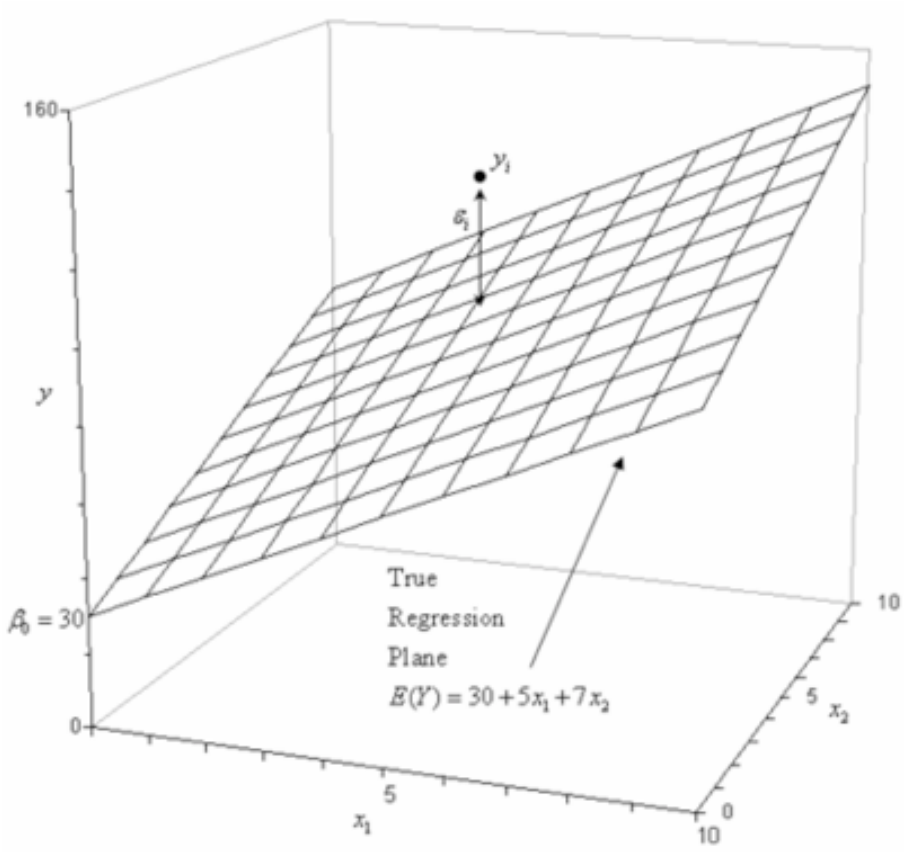
Intercept, värdet när alla ov i modellen är 0

Lutning: Ökning av y, när x ökar om en enhet och alla andra variabler hålls konstant för hela urvalet

Lutning: Ökning av y, när x ökar om en enhet och alla andra variabler hålls konstant för hela urvalet…

Multipel regression

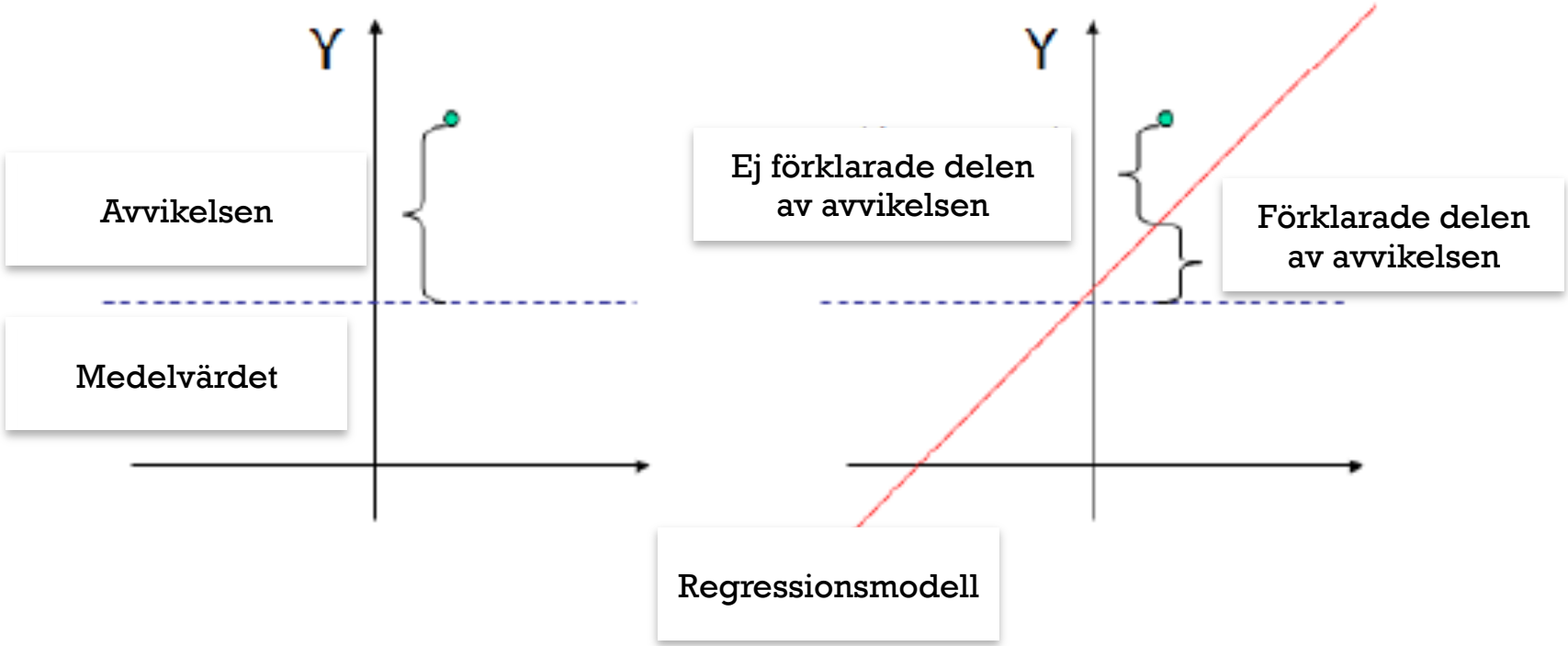
- Samma logik som enkla regressionen, dock mycket mer komplex att synliggöra



Hur väl fungerar vår modell? R^2

- …säger hur mycket av variansen (i procent) av vår beroende variabel y (avvikelse från medelvärdet) kan förklaras genom våra oberoende variabler
- → säger, hur väl vår modell (vår förutsägelse, gissning) fungerar

Variansförklaring



Att justera R^2

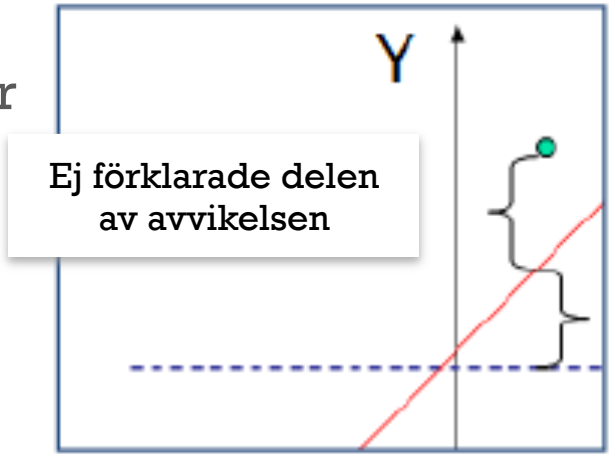
- Genom att ta in hur många variabler som helst in i modellen kan man
- Ju mer variabler, ju större R^2
- Man justera matematiskt i relation till storleken av urvalet
- Ju större N ju mindre faran att en ojusterat R^2 anger för stora värde
- SPSS/PSPP kan stegvis ta in flera variabler och räknar ut hur R^2 förändrar sig

Signifikans av modellen

- **T test** provar var en variabels oberoende statistiskt signifikans
  - Nollhypotesen = Den enstaka oberoende variabeln har inte någon påverkan på beroende variabeln utanför detta urvalet/ Påverkan är slumpmässig
  - Alternativhypotesen = Den enstaka oberoende variabeln har någon påverkan på beroende variabeln utanför detta urvalet
- **F test** provar hela modellens signifikans
  - Nollhypotesen = alla i modellen ingående oberoende variabler tillsammans (dvs. hela modellen) har har inte något påverkan på beroende variabeln/påverkan är slumpmässig
  - Alternativhypotesen= alla i modellen ingående oberoende variabler tillsammans (dvs. hela modellen) har har inte något påverkan på beroende variabeln

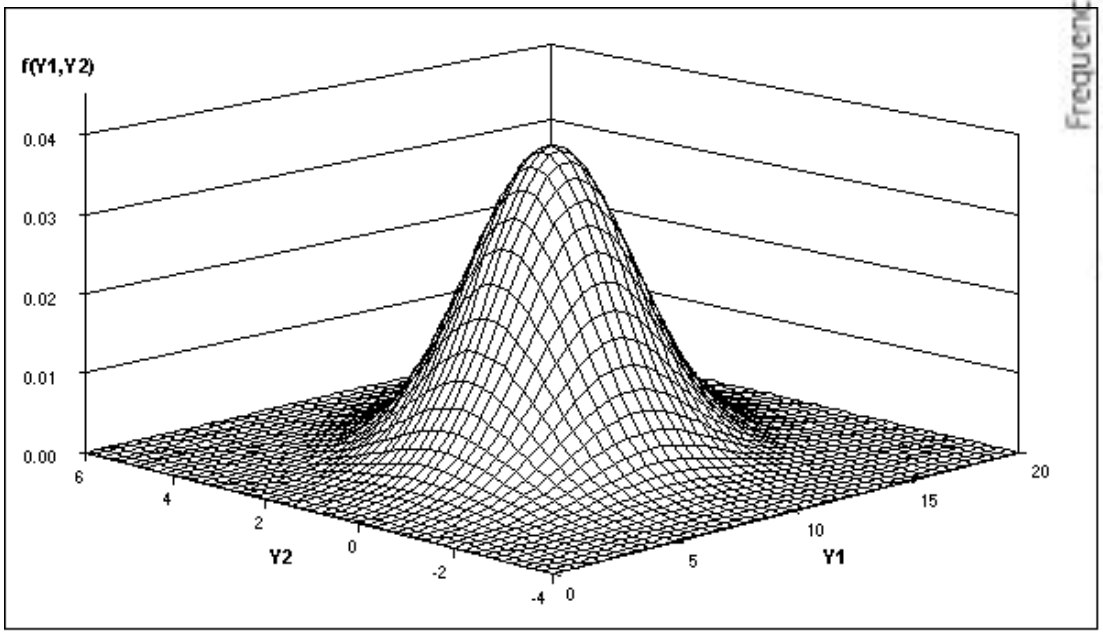
Standardfel av modellen

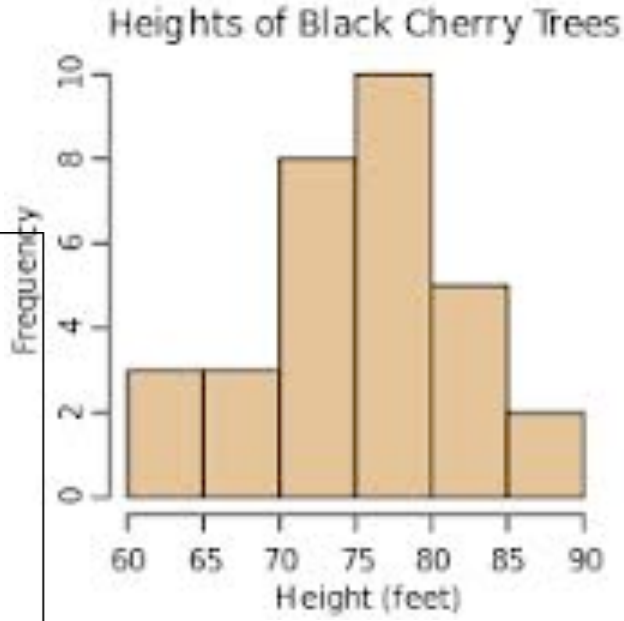
- Detta är standardavvikelsen av vår fel, dvs. detta som inte förklaras av modellen (typ motsatsen av $R^{2)}$
- Beskriver spridningen av våra y värde (BV) kring vår regressionslinje
- Anger hur mycket de ”sanna” värde avviker från vårt modell
- Ju mindre värdet, ju bättre vårt förutsägelse
- Bra för att jämföra olika modeller med varandra



Förutsättningar för multipla regressioner

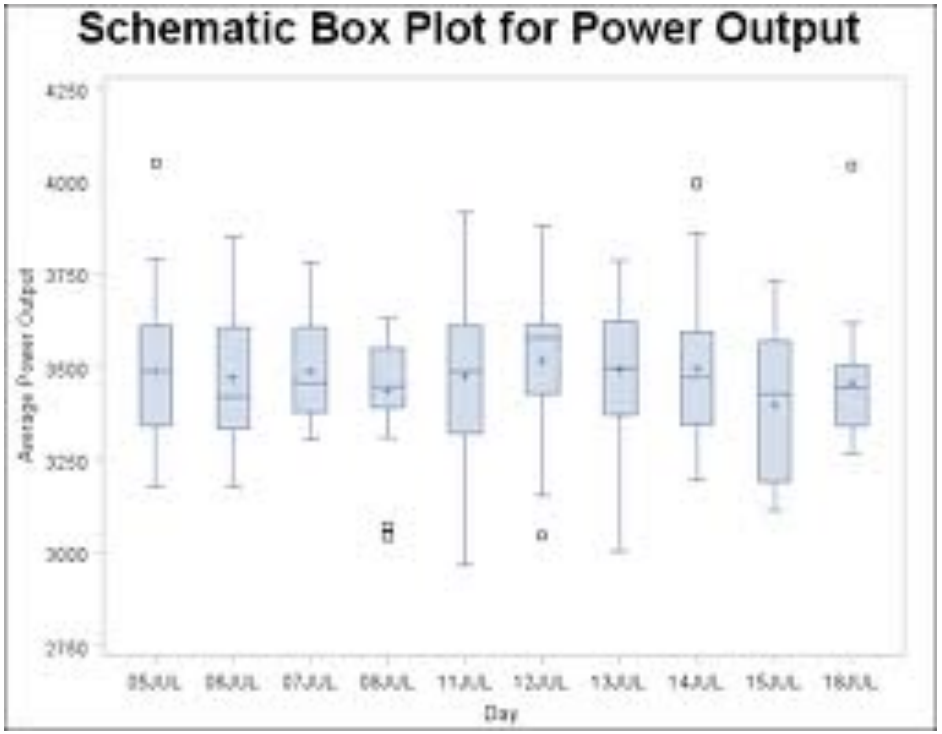
- Normalfördelning av alla variabler
- Prova man bäst med histogram
- Gäller inte dikotoma variabler (dummys)





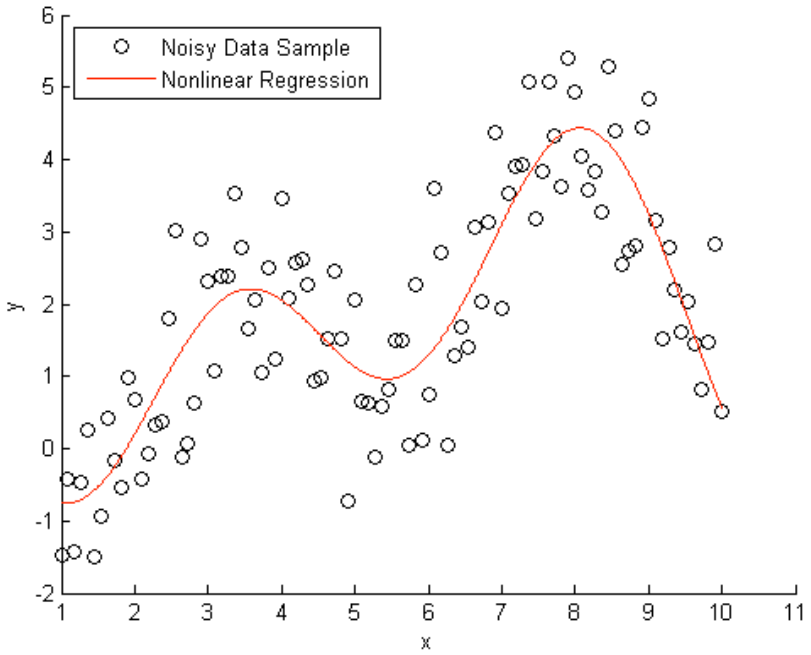
Förutsättningar

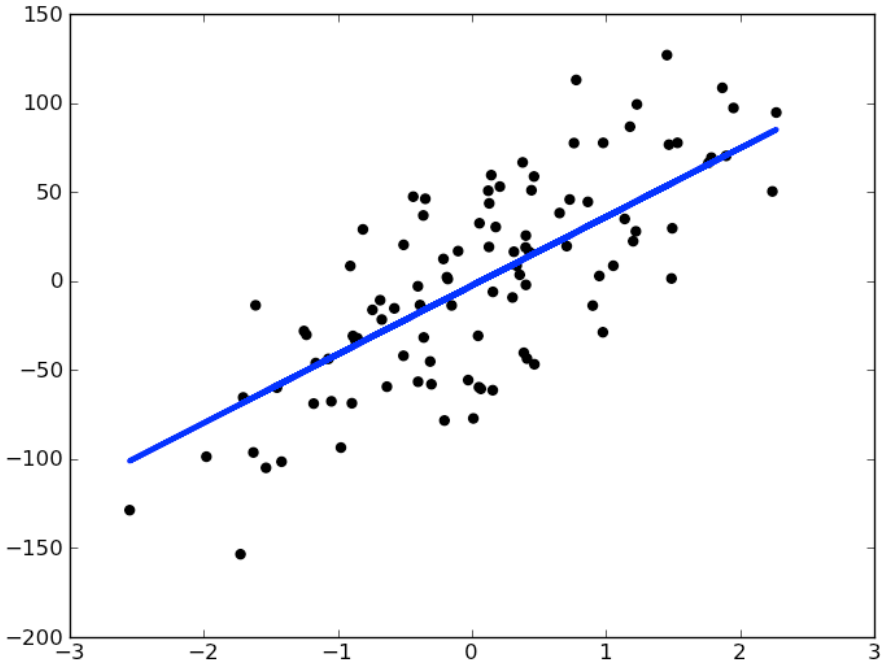
- **Inga extrema uteliggare**
- Testar man med **boxplotdiagram**



Förutsättningar

- Sambandet är linjärt
- Testa man bäst med scatterplot (för var en variabel)

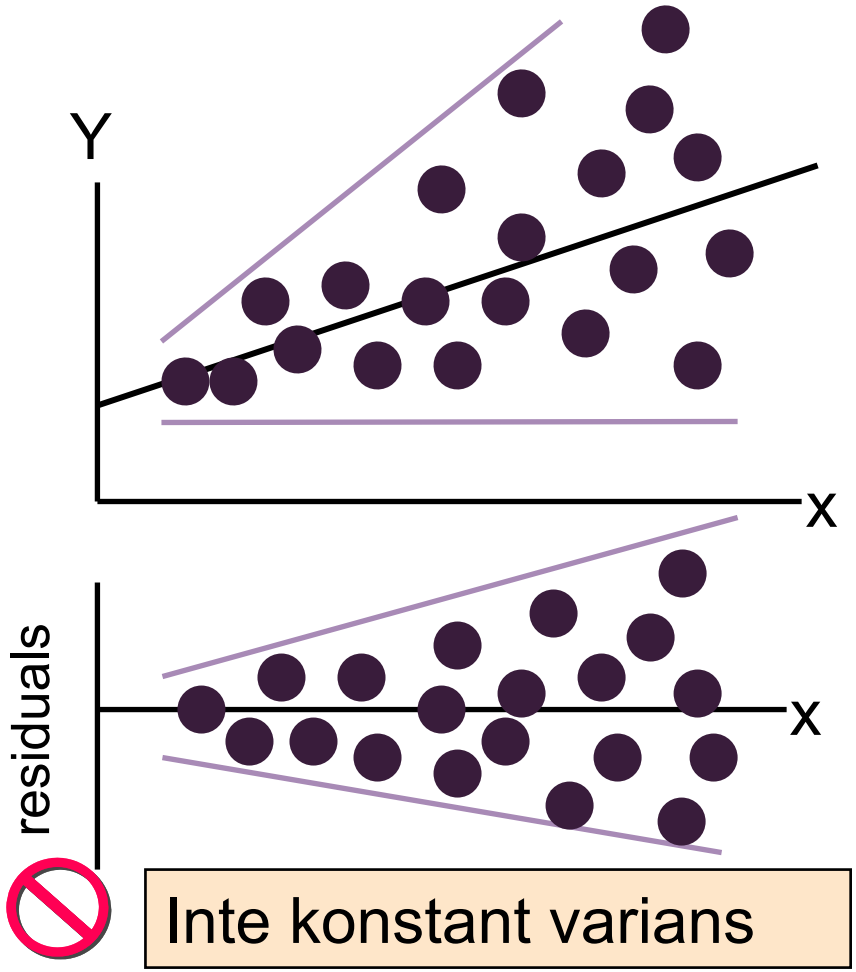


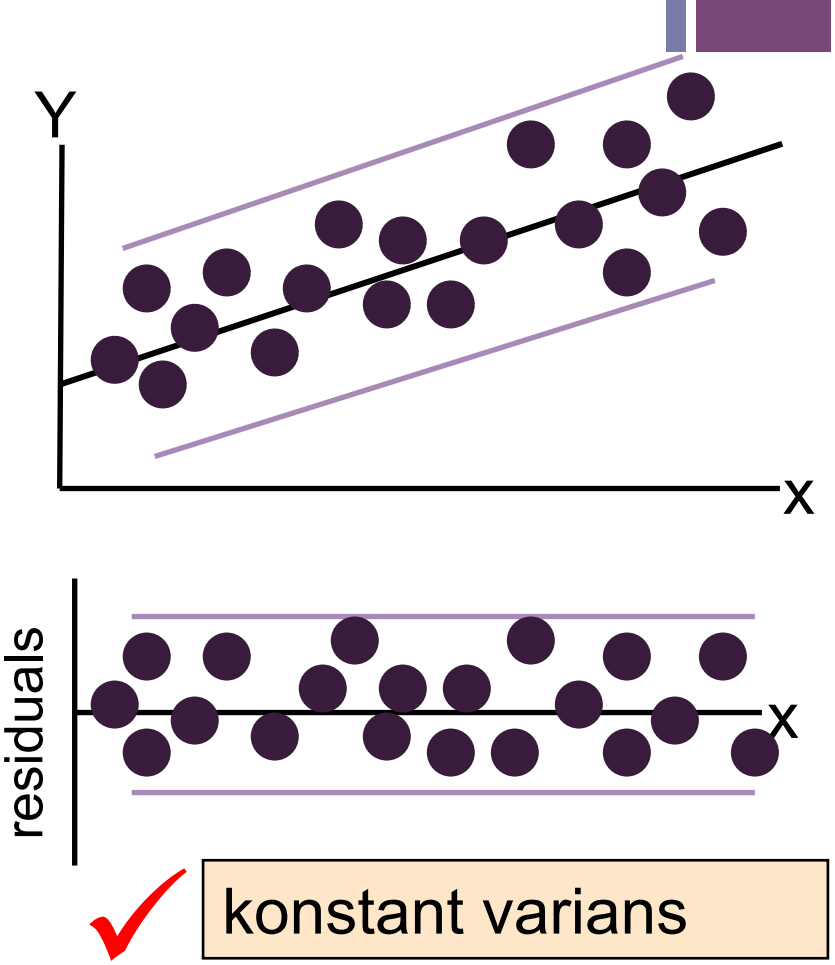


Förutsättningar: **felet (residual)**

- Felet, dvs. delen av data som inte kan förklarats genom modellen (**residual**) ska uppfylla flera olika krav för att vårt modell ska ge pålitliga resultat (**SPSS/PSPP använder sig av olika tester**)
- **Homoskedastitet**
  - Alla residualer har samma avstånd från vår modellinje
- Residualer ska vara normalfördelade
- Ingen **autokorrelation**
  - Det finns ingen tydligt samband mellan feltermen
- Ingen **multikollinaritet**
  - Alla oberoende variabler är (hyfsat) oberoende från varandra

Residualanalys för homoscedasticitet

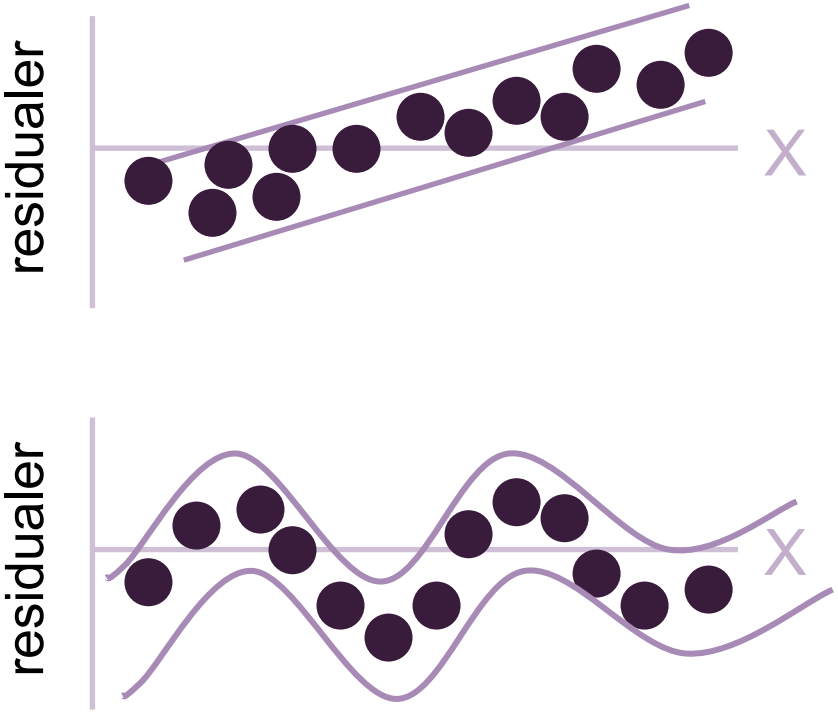




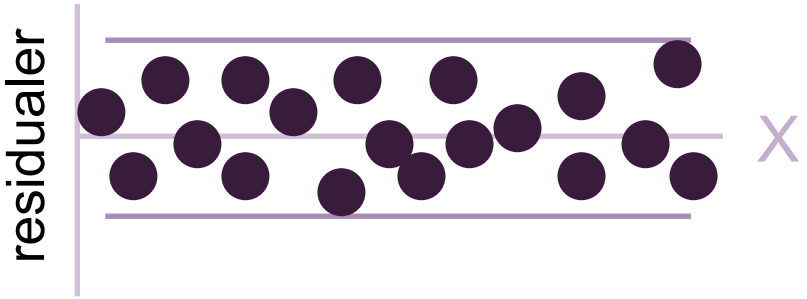
- källa: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Residualanalys för autokorrelation

beroende



oberoende

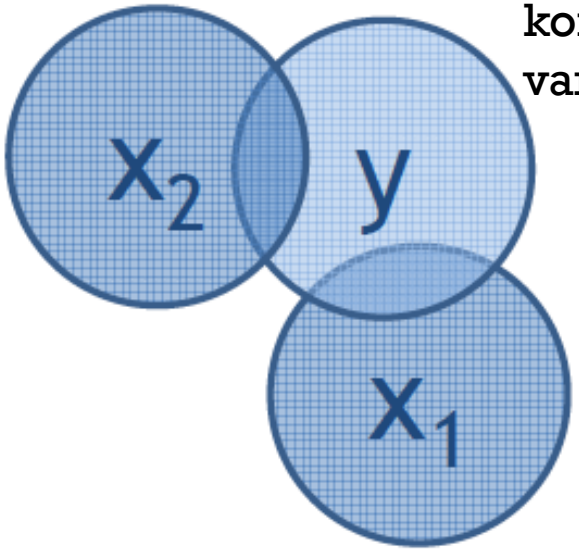


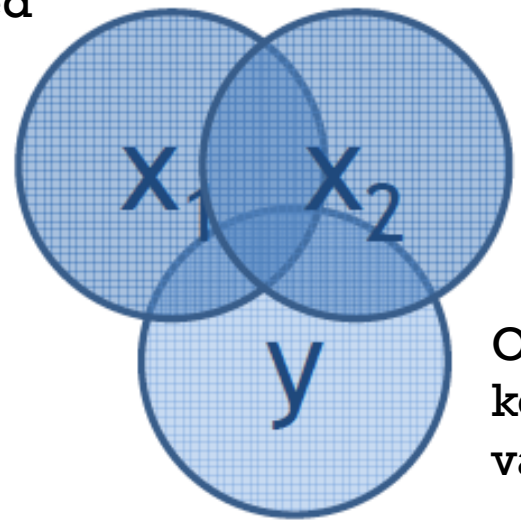
Kan även testas med **Durban-Watson** test: värde mellan 0 och 4
→ Värde runt 2: OK
→ Värde runt 0 eller 4: ej OK

Källa : Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Multikollinjaritet

Oberoende variabler korrelerar inte med varandra





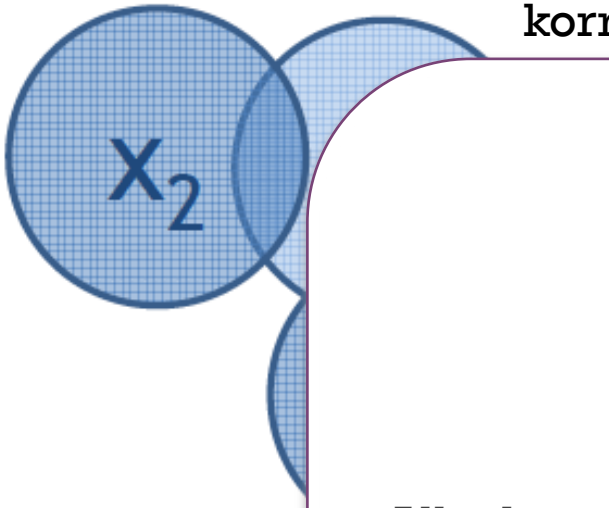
Oberoende variabler korrelerar starkt med varandra

Ju större multikollinjaritet, ju större blir standardfelet, vårt resulat blir oprecist
Två värde (som SPSS/PSPP räknar ut):

→ ***Tolerance***: hur mycket relation mellan OV inte påverkar resultatet (borde inte gå under 0,25)
→ **Variance inflation factor** (VIF): vid samma relaterade variabler blir varians dubbelt (1= ingen relation, över 5: ej OK)

Multikollinjaritet

Oberoende variabler
korrelerar inte med



Alla dessa procedurer kan hjälpa oss även i kvalitativ forskning för att undersöka huruvida våra (kvalitativa) modeller stämmer

Ju större r
Två värde

→ ***Tolera***
inte gå
→ **Varian**
varians d

Regressioner med kategoriala variabler: *Dummy* variabler

- Kan bara ta två värde: 0, 1
  - Kön, någon kriterium finns/finns inte (t.ex. universitetsexamen(ej universitetsexamen) osv.

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 D_2$$

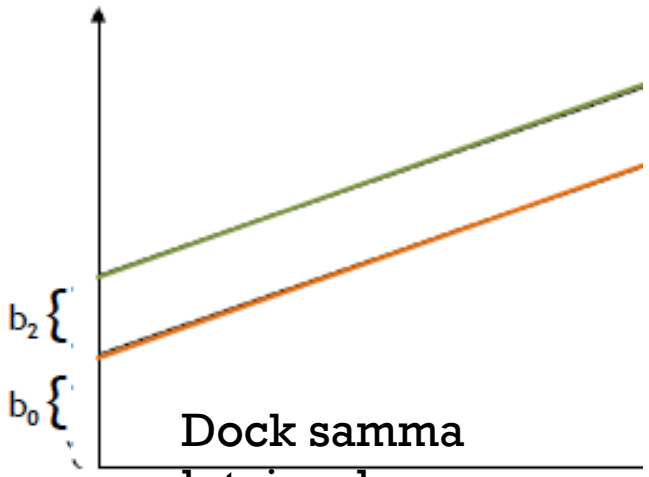
Dummyvariabel

D=0

$$\hat{y}_i = b_0 + b_1 x_1$$

D=1

$$\hat{y}_i = b_0 + b_1 x_1 + b_2$$



Dock samma lutning, kan undersökas med interaktionstermer

Flera än två kategorier?

- Har vi mer än två kategorier i i vår kategorial variabel, så skaffa vi flera dummy variabler
  - **Dock alltid en mindre än vi har kategorier!**

$D_1 =$ { 1 om bilen är vit
0 om bilen ej är vit

$D_2 =$ { 1 om bilen är silverfärgad
0 om bilen ej är silverfärgad

Kategorin "Andra färger" får värdena:
$D_1 = 0; D_2 = 0$

Exempel (sälja en bil)

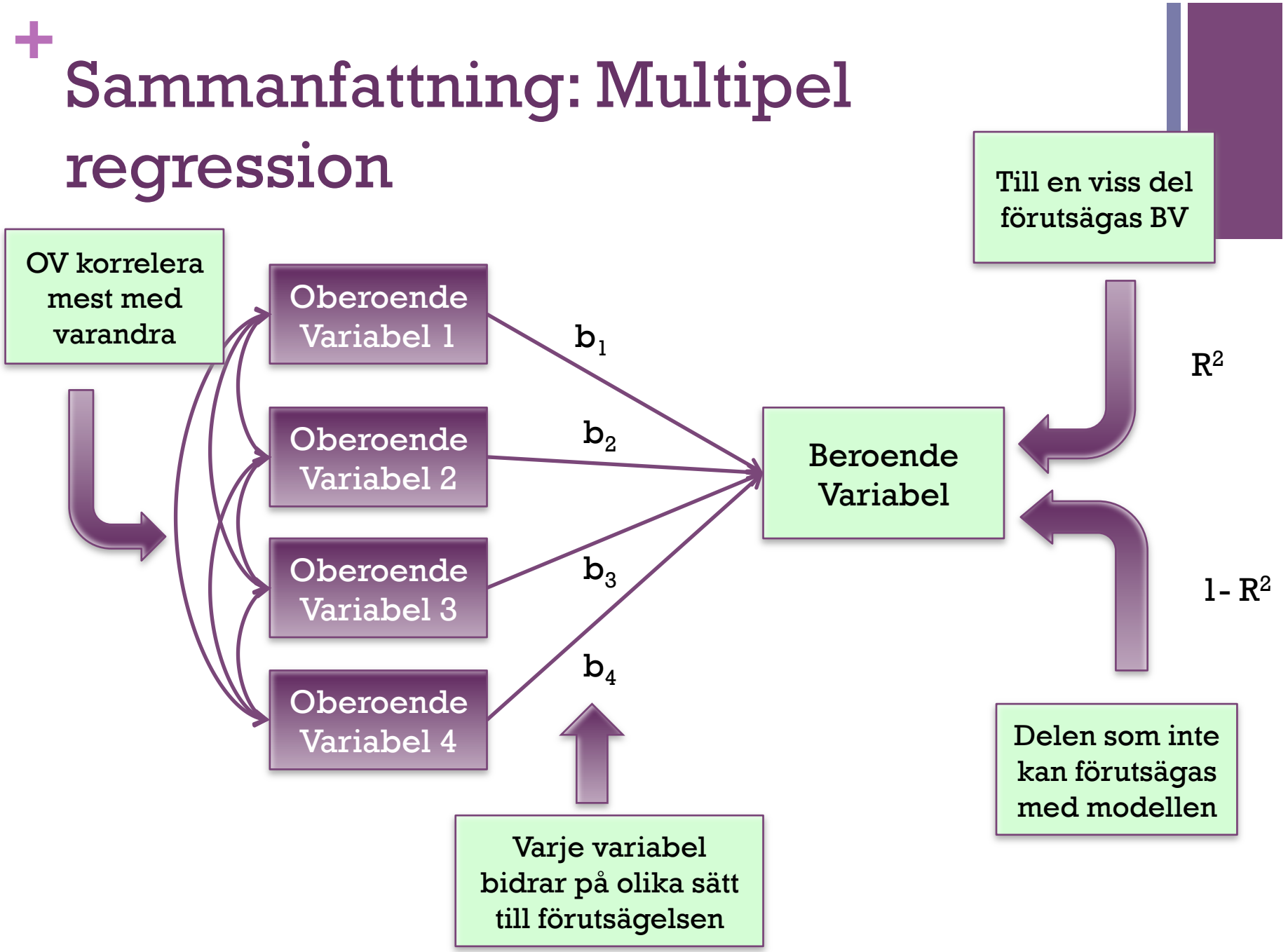
$$\text{Pris (bil)} = 16701 - .0555(\text{mätare}) + 90.48(D_1) + 295.48(D_2)$$

För ytterligare en mil minskar priset med i genomsnitt 5.55 cents.

En vit bil säljs, i genomsnitt, För 90.48 Euro mer än en bil i kategorin "annan färg"

En silverfärgad bil säljs, i genomsnitt för 295.48 Euro mer än en bil i kategori "annan färg"

Sammanfattning: Multipel regression



Exempel

	b	SE *b*	*β*	t	signifikans
(Constant)	4.202	.451		-9.314	.000
Summa talserier	.231	.010	.447	23.481	.000
Summa motsatser	.152	.014	.202	11.237	.000
Kön	.016	.129	.002	.124	.901
Bakgrund	.891	.276	.081	3.234	.001

Bakgrund: 1=svensk/0=inte svensk
Kön: 1=kvinnor/0=män

Tabell 3. Regressionsekvationen för den beroende variabeln Total poäng på matematikprov i åk6 (Källa Carina Carlhed)