

BI1363 HT 2020

Korrelation och regression

Oktober 2020

Adam Flöhr, BT, SLU

Regressionsanalys och korrelation

Motsvarar *Biometri*, kap 11

I korthet

Vi samlar in två numeriska variabler från en samling objekt

Vill undersöka hur de förhåller sig till varandra

Med **regression** kan vi skapa en modell för variabel som en funktion av den andra

Med **korrelation** kan vi mäta styrkan på sambandet mellan variablerna

Sambandet mellan variablerna kan testas med **t-test** eller **F-test**

Korrelation och regression

Korrelation och regression är metoder för att mäta ett samband mellan två eller flera variabler

Regression

En variabel (en *svarsvariabel*) förklaras av en andra variabel (en *förklarande variabel*)

För ett givet värde av den förklarande variabeln, vilket är det förväntade värdet av svarsvariabeln?

Korrelation

Graden av linjärt förhållande mellan två numeriska variabler

De två variablerna är av *lika* betydelse (du kan byta deras ordning och korrelationen är densamma)

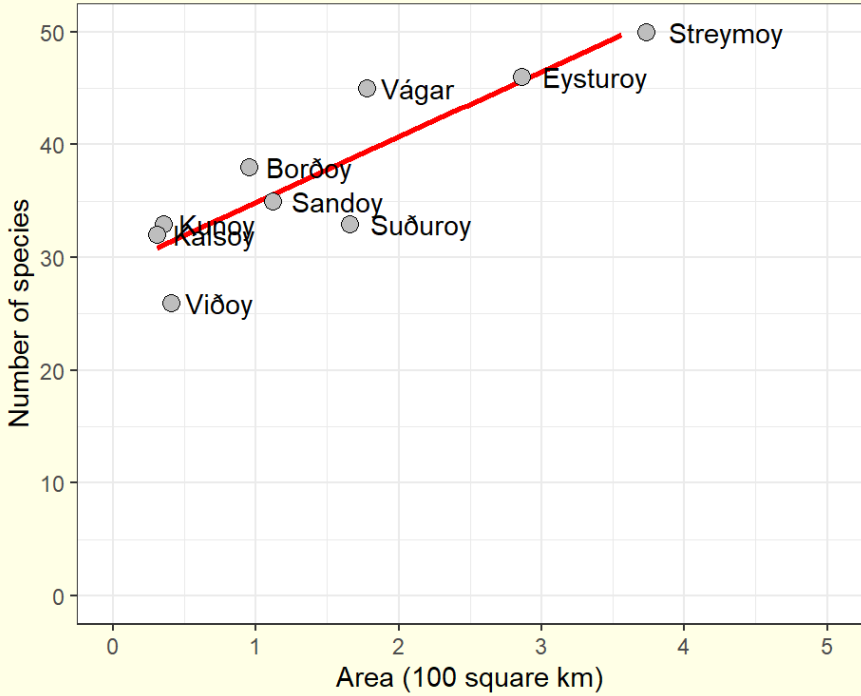
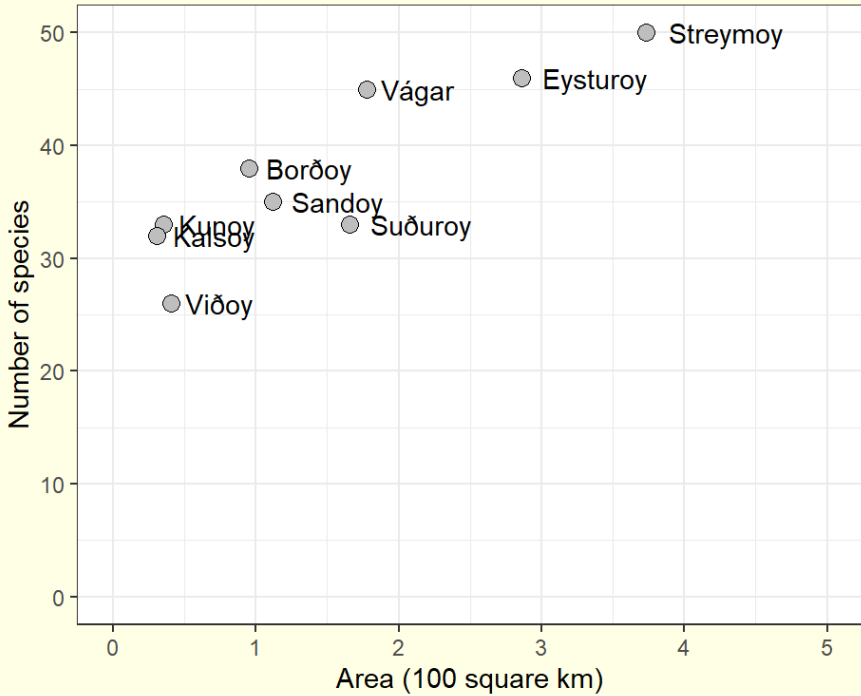
Enkel linjär regression

Två kontinuerliga variabler, x och y

Parade så att x_i och y_i kommer från samma objekt

Modell för y som en funktion av x

Name	Area (100 square km)	Number of species
Streymoy	3.735	50
Eysturoy	2.863	46
Vágar	1.776	45
Suðuroy	1.660	33
Sandoy	1.121	35
Borðoy	0.950	38
Viðoy	0.410	26
Kunoy	0.355	33
Kalsoy	0.309	32



Modellformulering

För en observation av y kan vi formulera följande modell

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad e_i \sim N(0, \sigma^2)$$

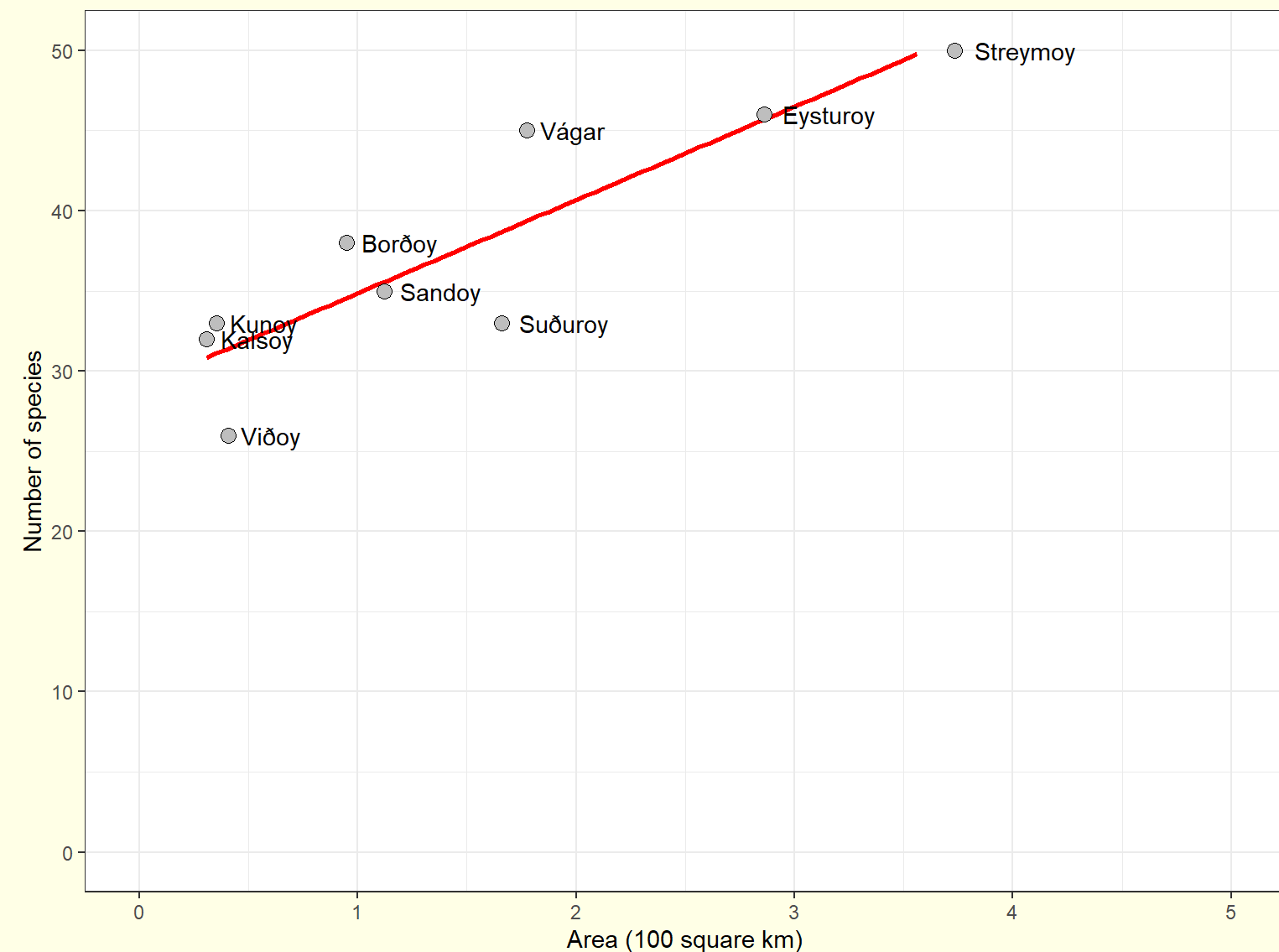
β_0 - intercept, det uppskattade värdet av y när x är noll

β_1 - lutning, den beräknade marginella förändringen i y när x ökar med ett

Residualer: $\hat{e}_i = y_i - \beta_0 - \beta_1 x_i$

Variabeln y kallas *svarsvariabel*, *responsvariabel*, *beroende variabel* eller *förklarad variabel*

Variabeln x kallas *oberoende variabel* eller *förklarande variabel*

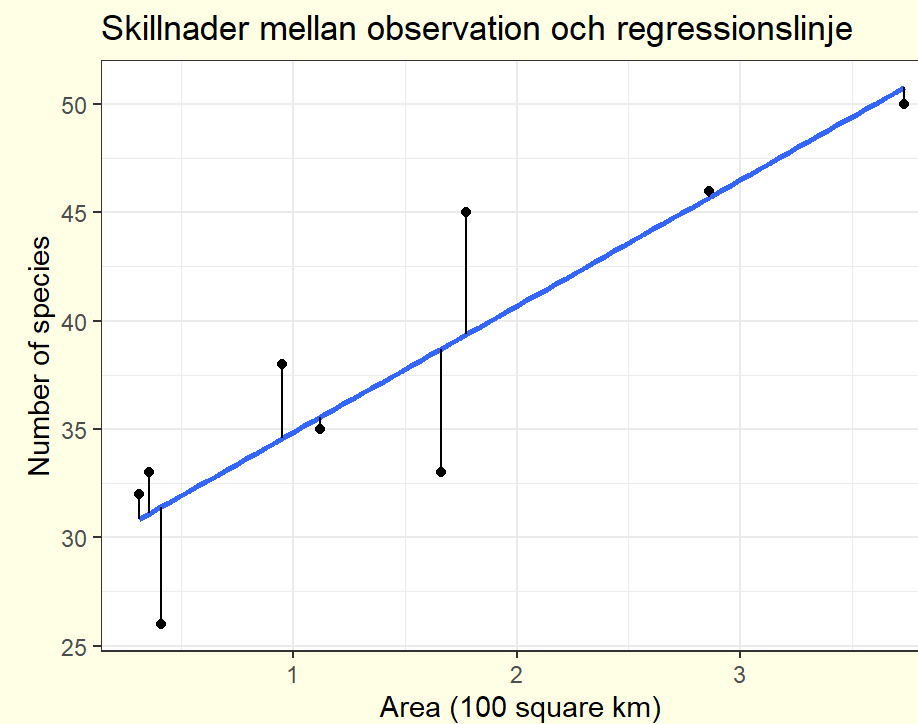
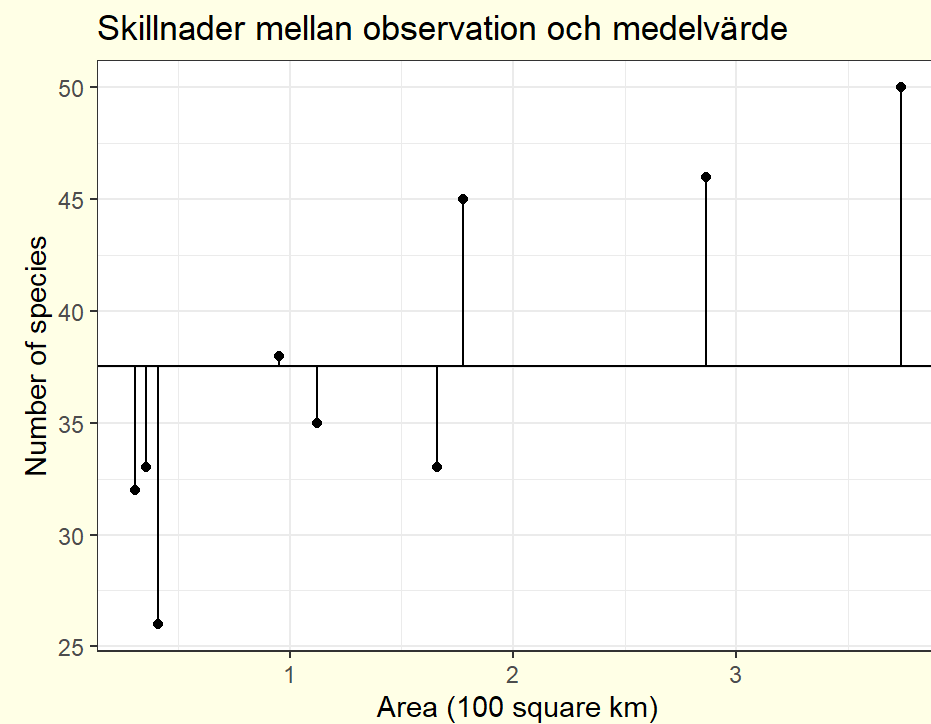


Skattning av intercept och lutning

Om $\hat{\beta}_0$ och $\hat{\beta}_1$ är skattningar av intercept och lutning, ges residualen av

$$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

De skattade parametrarna $\hat{\beta}_0$ och $\hat{\beta}_1$ väljs så att summan av kvadrerade residualer minimeras



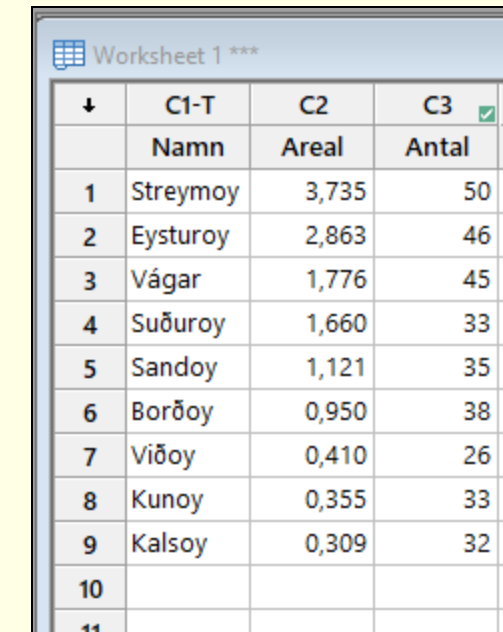
Minitab-utskrifter

Data skrivs in i ett *Worksheet* med en rad för varje observation

Regressionen skattas genom *Stat > Regression > Regression > Fit Regression Model...*

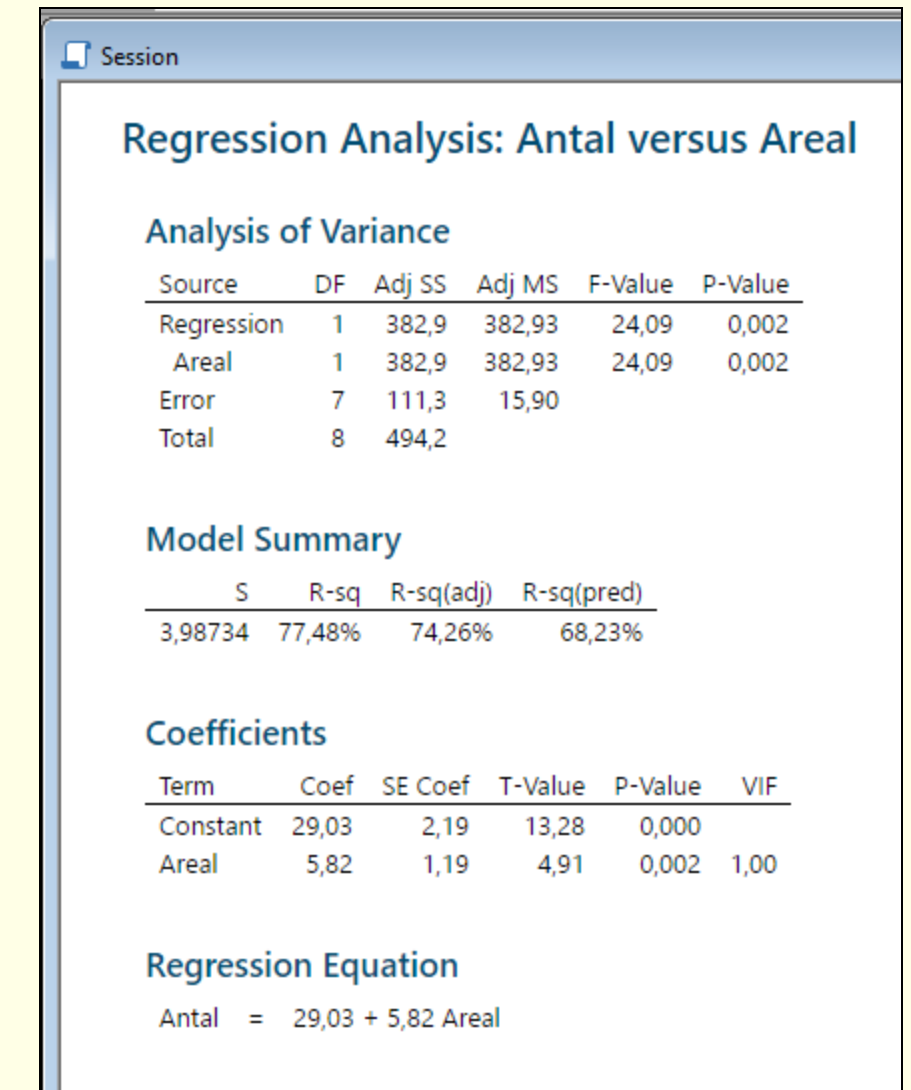
Utskriften ger

- *Analysis of Variance*. En anova-tabell lik den vi sett tidigare
- *Model summary*. Mått på modellanpassning
- *Coefficient*. Skattade parametrar
- *Regression equation*. Den skattade regressionsekvationen



Worksheet 1 ***

↓	C1-T	C2	C3	
	Namn	Areal	Antal	
1	Streymoy	3,735	50	
2	Eysturoy	2,863	46	
3	Vágar	1,776	45	
4	Suðuroy	1,660	33	
5	Sandoy	1,121	35	
6	Borðoy	0,950	38	
7	Viðoy	0,410	26	
8	Kunoy	0,355	33	
9	Kalsoy	0,309	32	
10				
11				



Session

Regression Analysis: Antal versus Areal

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	382,9	382,93	24,09	0,002
Areal	1	382,9	382,93	24,09	0,002
Error	7	111,3	15,90		
Total	8	494,2			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3,98734	77,48%	74,26%	68,23%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	29,03	2,19	13,28	0,000	
Areal	5,82	1,19	4,91	0,002	1,00

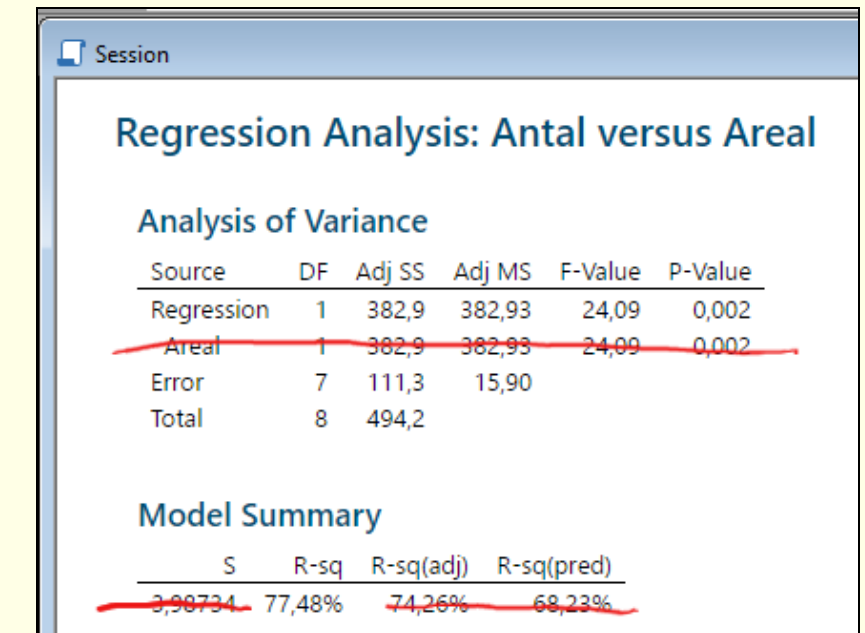
Regression Equation

Antal = 29,03 + 5,82 Areal

Anova-tabell och modellanpassning

Tolkningen av anova-tabellen liknar den vid variansanalys

- *DF*. Antal frihetsgrader. Regressionen har en frihetsgrad, residualen har $N - 2$ frihetsgrader, och totalen $N - 1$ frihetsgrad
- *Adj SS*. Kvadratsummor beräknade från datan
- *Adj MS*. Kvadratsumman delat på antalet frihetsgrader
- *F-Value*. Testvärde i ett F-test. Ges av $\frac{MS_R}{MS_e} = \frac{382.93}{15.90} = 24.09$
- *P-Value*. P-värde för F-testet beräknat som svansen i en F-fördelning med $\nu_1 = 1$ och $\nu_2 = N - 2$



Analysis of Variance					
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	382,9	382,93	24,09	0,002
Areal	1	382,9	382,93	24,09	0,002
Error	7	111,3	15,90		
Total	8	494,2			

S	R-sq	R-sq(adj)	R-sq(pred)
3,98734	77,48%	74,26%	68,23%

Determinationskoefficienten

Det mest relevanta måttet i *Model Summary* är *determinationskoefficienten* R^2 (R-sq)

R^2 ges av kvoten SS_R/SS_T och mäter hur mycket av den totala variationen som förklaras av regressionsmodellen

Skattade parameterar och modellekvation

Coefficient ger skattningarna av modellens parameterar

Interceptet står som *Constant* och lutningen som den oberoende variabeln (här *Areal*)

- *Coef.* Själva skattningen
- *SE Coef.* Skattningens medelfel
- *T-Value.* Skattningen delat på medelfelet. Testvärde vid ett t-test
- *P-value.* P-värdet beräknat från en t-fördelning med $N - 2$ frihetsgrader
- *VIF.* Ej relevant vid enkel linjär regression

Regression Equation ger den skattade modellekvationen

En beräkning för ett visst x-värde ger en prediktion av y-variabeln vid det x-värdet

Det predikerade antalet arter på en 200 kvadratkilometer stor ö väntas vara

$$29.03 + 5.82 \cdot 2 = 40.67$$

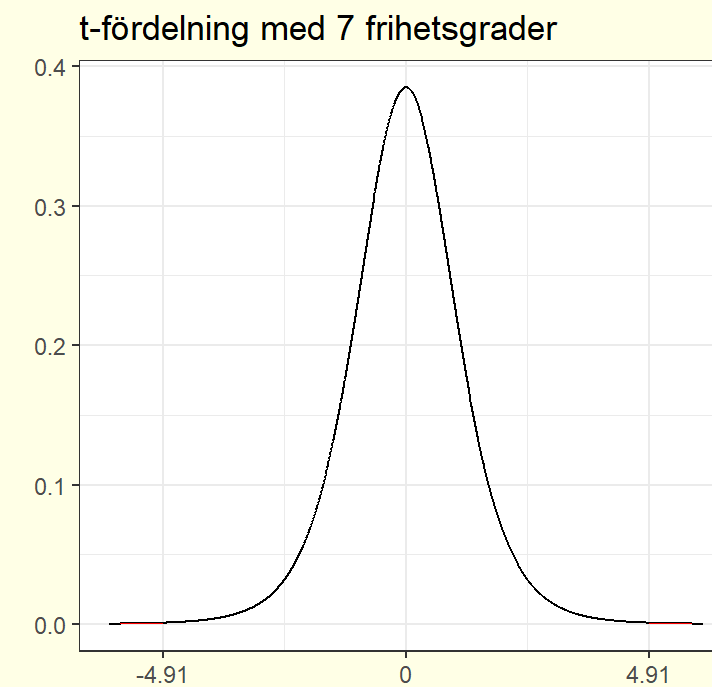
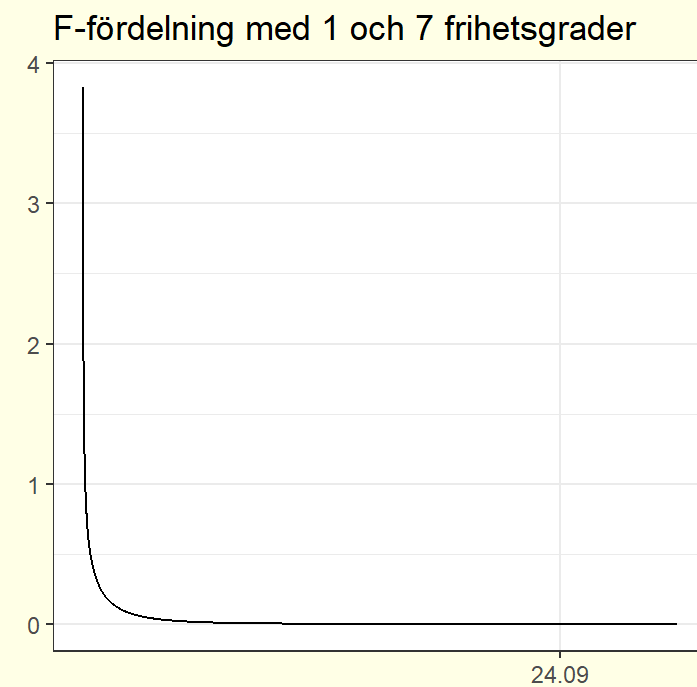
Coefficients					
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	29,03	2,19	13,28	0,000	
Areal	5,82	1,19	4,91	0,002	1,00
Regression Equation					
Antal = 29,03 + 5,82 Areal					

Tester

Vi kan göra hypotestester på β_0 och β_1

Det vanligaste testet är av $H_0 : \beta_1 = 0$, dvs om det finns en lutning skild från noll

Detta kan testas med ett F-test eller ett t-test



Antalet frihetsgrader för t-fördelning och för nämnaren i F-fördelningen ges av $N - 2$

Modellantaganden

Statistiska tester av regressionsmodellen bygger på följande antaganden

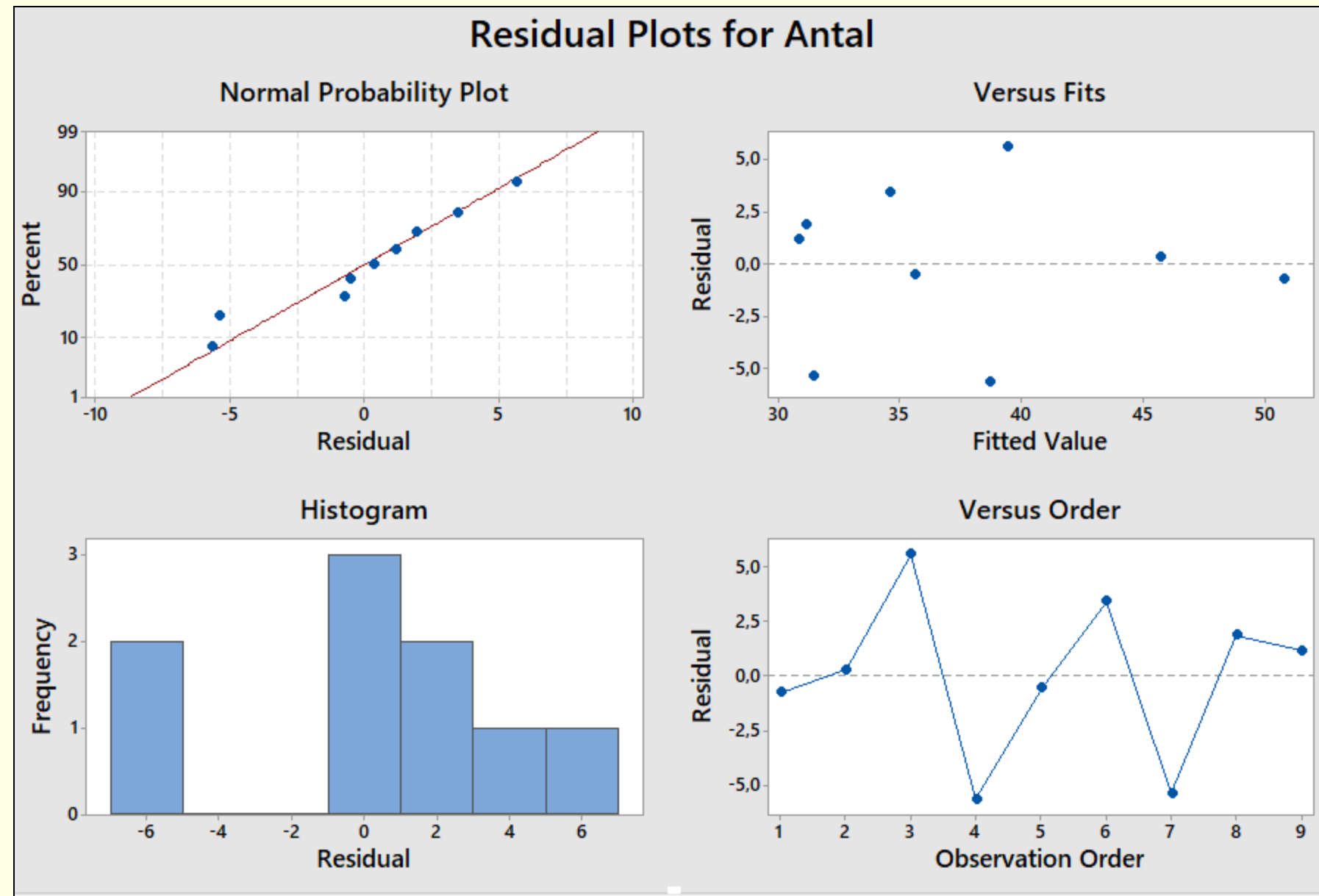
- Normalfördelade residualer
- Lika varians i y för alla nivåer av x
- Oberoende observationer

Antaganden kan testas genom formella tester eller genom residualplottar

Residualplottar

Minitab ger fyra grundläggande residualplottar

(Klicka i *Four in one* under *Graphs* i regressionsfönstret)



Normalfördelningsantagandet

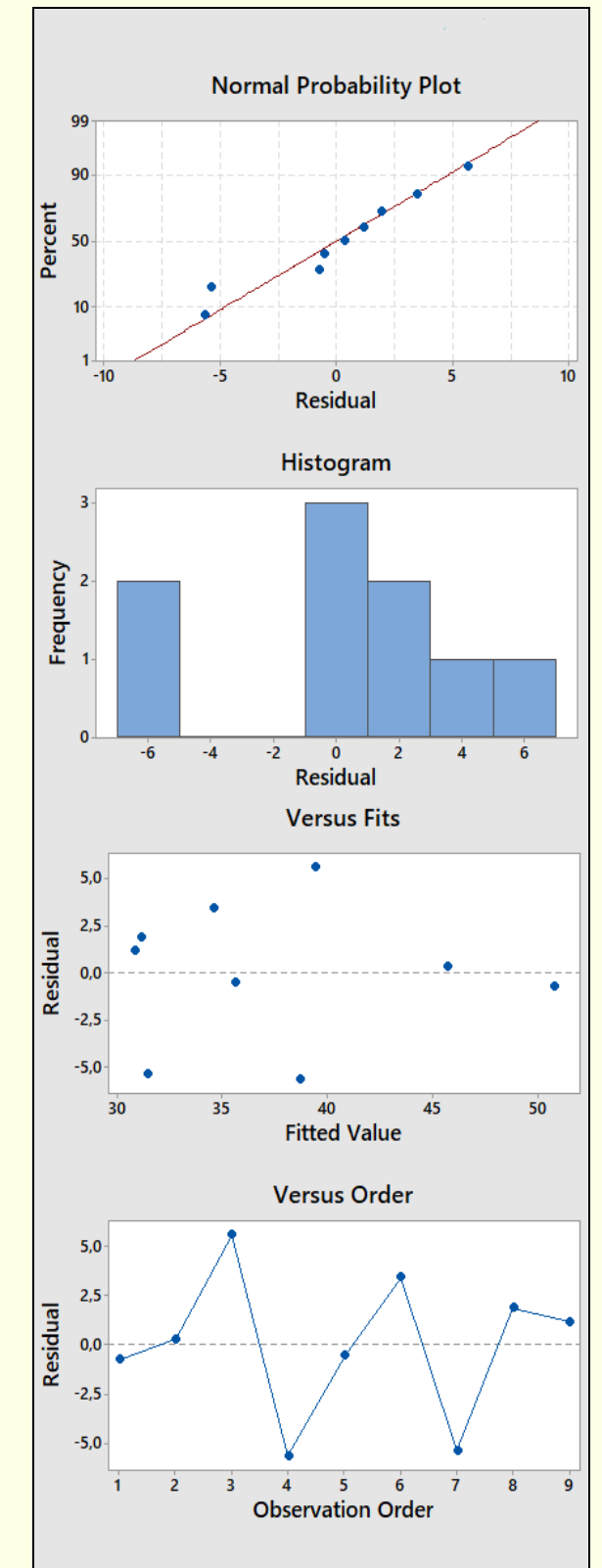
- I normalfördelningsplotten (övre vänster) ska punkterna ligga nära den diagonala linjen
- I histogrammet ska observationerna följa en normalfördelning

Lika varians

- I spridningsdiagrammet (övre höger) ska spridningen kring den streckade linjen vara densamma för olika x-värden

Oberoende

- I linjediagrammet (nedre höger) ska det saknas mönster - residualerna ska vara oberoende av observationernas ordning
- Bara relevant om observationerna är ordnade i någon naturlig ordning, t.ex. om man samlat in data över tid



Hantering av brister i antaganden

Vanligt att åtgärda brister i modellantaganden

Datatransformationer

- Istället för y , skattas modellen på \sqrt{y} eller $\log(y)$
- Parametertolkning måste ske i ljuset av transformationen

Granskning av extremvärden

- Om det finns tydliga extremvärden kan man undersöka om det är resultatet av någon felmätning
- I vissa fall kan extremvärden uteslutas ur analysen

Icke-parametriska metoder

- Statistiska metoder som saknar fördelningsantagande
- Se *Biometri*, kap 12. (Ej del av kursen.)

Multipel regression

- Enkel regression kan utvecklas med *multipel regression*
- Medtar ytterligare variabler i modellen

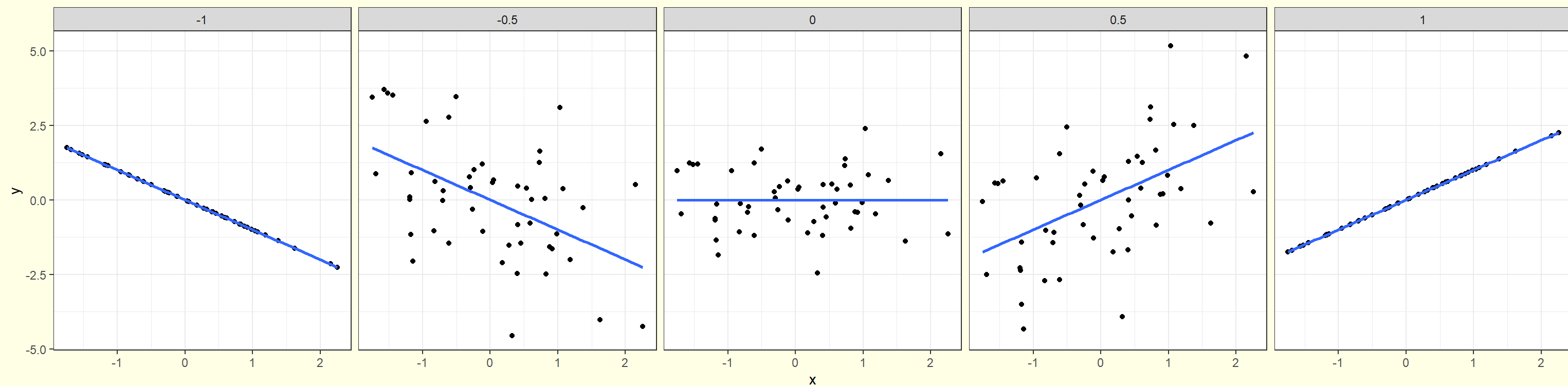
Korrelation

Sambandet mellan två numeriska variablerna kan också undersökas med korrelation

Korrelation är ett mått på graden på **linjär** relation mellan två variabler

Variablerna ses som likvärdiga - man ser inte den ena som en funktion av den andra

Anges som ett tal mellan -1 och 1, där negativa värden pekar på ett negativt samband och positiva på ett positivt samband



Beräkning

Det finns olika typer av korrelation, den vanligaste är *Pearsonkorrelation*, betecknad med r

Pearsonkorrelation beräknas genom

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / (n - 1)}{s_x \cdot s_y}$$

Täljaren $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / (n - 1)$ är *kovariansen*

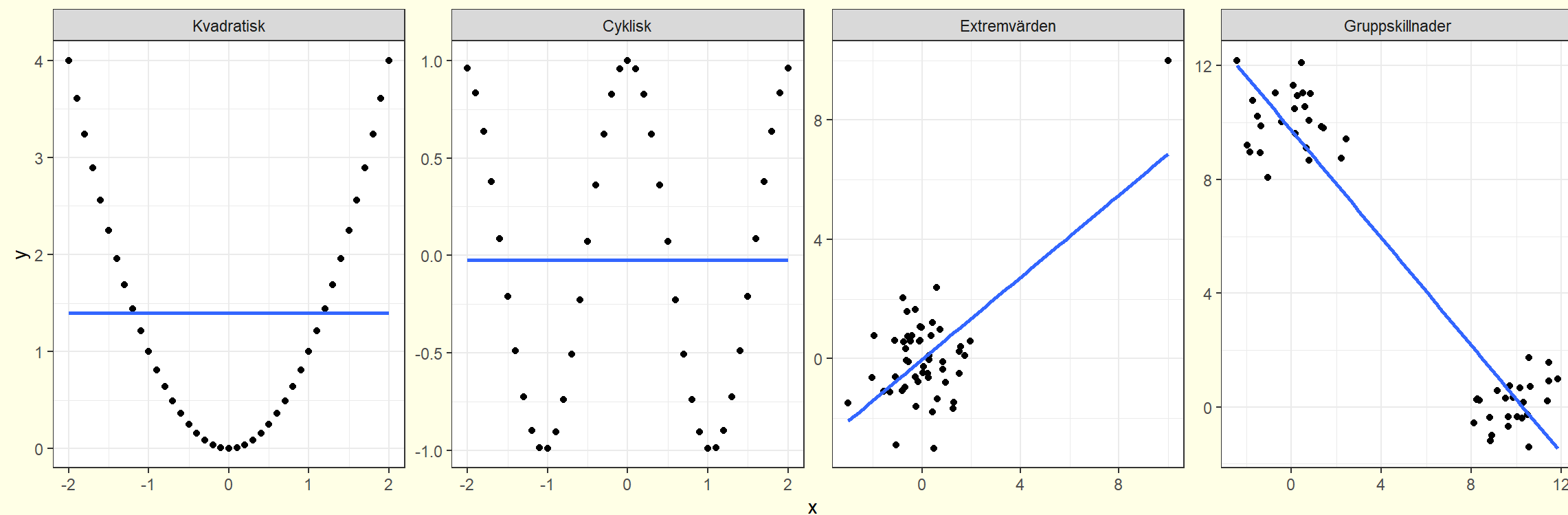
Linjär relation och oberoende

Korrelation mäter det *linjära* sambandet

Det kan finnas ett starkt samband mellan två variabler även om korrelationen är låg

Korrelationen kan vara hög på grund av extremvärden

Storskaliga skillnader mellan grupper kommer att påverka korrelationen mer än småskaliga skillnader inom grupper
(*Simpsons paradox*)



Korrelationerna är 0, 0, 0.66 respektive -0.95

Determinationskoefficienten och tester

Korrelation r i kvadrat ger determinationskoefficienten R^2 från regressionen

Korrelation kan testas med nollhypotesen $H_0 : r = 0$ (okorrelerade variabler) genom ett t-test

Testet ger samma utfall som ett test för $H_0 : \beta_1 = 0$ i en regression

Slut