Version, 01-09-05

Institutionen för marin ekologi, Göteborgs universitet

STATISTISK ANALYS OCH UTFORMNING AV EKOLOGISKA EXPERIMENT
2. STATISTISKA MODELLER OCH REGRESSION

Sammanställt av Per Jonsson 1995, modifierat av Mats Lindegarth 2001, Bo Johannesson 2001

Modellbegreppet

Efter introduktionen av fördelningar, hypotestester och statistiska fel, möter vi här några begrepp och tankemönster som vi kommer att få mycket glädje av framöver. Som vi har sett (i kompendium 1) visar observationer (data) variation som vi bl.a. såg hos populationer av blåmussla. Gör tankeexperimentet att vi fått ett stickprov från ett salutorg med blåmusslor från både ost- och västkusten. Det kan t.ex. se ut som i Figur 1.

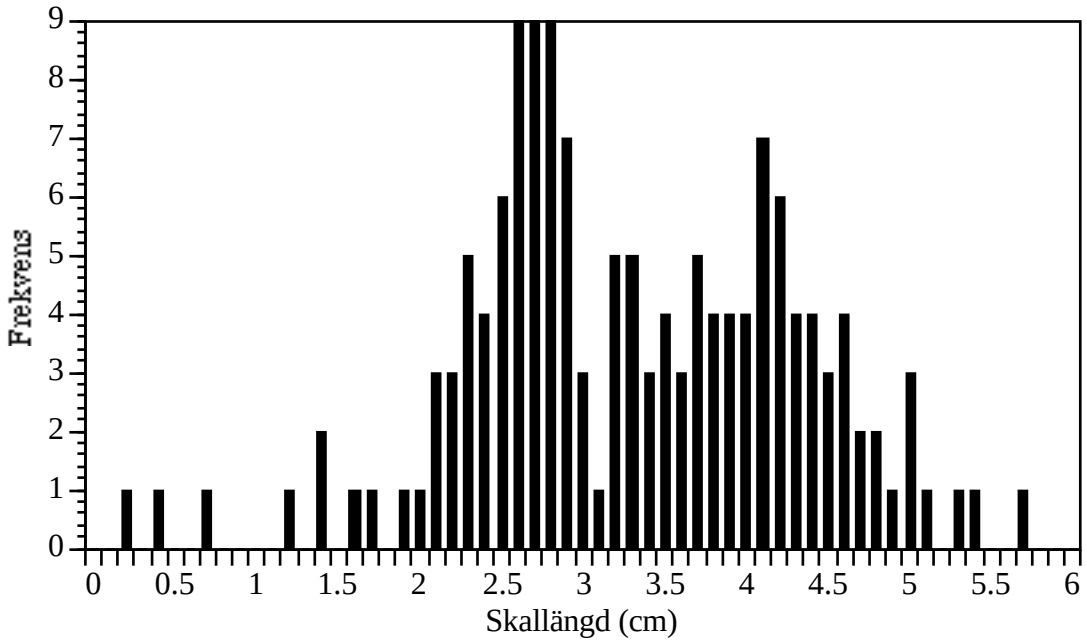

Fig. 1 Frekvensfördelning av musslors längd.

Man säger ofta att man sedan statistiskt analyserar sina data. Vad man egentligen gör är att ställa upp en s.k. statistisk modell och ser hur väl modellen förklarar variationen i observationerna. Vi kan se på det totala materialet och beräkna ett medelvärde, summa av kvadrerade avvikelser (SS = sum of squares) och varians: $\bar{x}$ = 3.3 cm, SS = 149, s^2 = 1.03 (s = 1.01). Vi kan skriva längden av varje mussla (L) i populationen som summan av det gemensamma medelvärdet (μ) och en individuell avvikelse (e_i). Subskriptet i är ett index för varje mussla, i = 1, 2, 3…n.

$L_i = \mu + e_i$

Vi kan nu utöka modellen för att försöka förklara ytterligare en del av variationen. En uppenbar modell är, verbalt uttryckt, att storleken på musslorna beror på om de kommer från ost- eller västkusten. D.v.s. en del av variationen i materialet kan förklaras med modellen.

$L_{ij} = \mu + Kust_i + e_{ij}$

Kust är i detta fall en s.k. faktor som ger ett bidrag till längden beroende på var musslorna kommer ifrån. Genom en genetisk analys kan vi identifiera de musslor som kommer från

respektive kust och separata medelvärden beräknas till 4.0 och 2.6 cm. Vi kan skriva modellen som:
$L_{ij} = 3.3 \pm 0.7_j + e_{ij}$
där ± 0.7 är effekten av Kust-faktorn. Hur mycket av den totala variationen återstår nu att förklara? De kvadrerade avvikelserna, SS, till respektive medelvärde blir:
36.4 + 33.4 = 69.8

Vi har alltså genom att utöka modellen ökat förklaringsgraden p.g.a. faktorn Kust med 44 % (149 - 69.8)/149. Det är nu möjligt att gå vidare med den kvarvarande variationen, den s.k. residual variationen, och kanske öka förklaringsgraden med nya faktorer, t.ex. ålder. Genom att räkna tillväxtzoner är det möjligt att dela in musslorna i tre åldersgrupper. För varje kustpopulation kan nu tre medelvärden räknas fram som representerar åldersgrupperna enligt modellen:
$L_{ijk} = \mu + Kust_i + Ålder_j + e_{ijk}$

Totalt förklaras nu 74 % av variationen. Residualvariation på 26 % är den oförklarade delen av den totala variationen och representerar alla okända faktorer som påverkar mussellängden t.ex. genotyp, födotillgång, temperatur etc. En uppdelning av variationen, i observationer eller experiment, på olika faktorer enligt en statistisk modell är ett mycket kraftfullt angreppssätt.

Faktorer

Vi ska titta lite närmare på olika typer av förklarande faktorer. En faktor är alltså en variabel som potentiellt förklarar (helst orsakar) den variation som vi observerar. En faktor kan variera och man talar om att den har olika nivåer, exempelvis:

Faktor	**Nivåer**
Ålder	1-åringar, 2-åringar
Årstid	vinter, vår, sommar, höst
Lokal	västkust, ostkust

Nu kommer ett något knepigt begrepp. Man delar nämligen in faktorer i två typer, fixerade (eng. fixed) och slumpade (eng. random). Om alla tänkbara nivåer av en faktor finns med i modellen är faktorn fixerad. Utgör däremot antalet nivåer som finns med ett slumpmässigt urval av alla tänkbara faktorer är den slumpad. Exempel på en fixerad faktor är årstid om vi mäter t.ex. tillväxt av gran under, vår, sommar, höst och vinter; alla tänkbara nivåer av årstid är med. Exempel på en slumpad faktor är om vi mäter

reproduktionen hos en ciliat vid de olika temperaturerna 8, 10, 17 och 23°C vilka representerar fyra slumpmässigt dragna temperaturer mellan 5 och 25°C. En god tumregel är att om man skulle välja samma nivåer om man gjorde om experimentet så är faktorn fixerad, annars slumpad. Skillnaden i betydelse mellan dessa två grupper av faktorer förstår man bara i steg så ha tålamod. Hur som helst är det viktigt att känna till om de faktorer man har är fixerade eller slumpade eftersom de statistiska analyserna i många fall är olika.

En första känsla av skillnaden mellan de två typerna av faktorer fås genom att se på exemplet med musslorna igen. Uppenbarligen är Kust i detta fall fixerad. Det finns naturligtvis fler kuster men i vår frågeställning är vi från början enbart intresserad av att jämföra ost- och västkusten i Sverige. Som vi såg förklarade denna faktor en hyfsad del av variationen. Den är dessutom s.k. statistiskt signifikant (H_0 förkastas) och man säger då att faktorn har en signifikant behandlingseffekt (eng. treatment effect). Beroende på nivån av Kust läggs en effekt till medelvärdet (μ) som i fallet med musslorna var ± 0.7. Så fungerar fixerade faktorer.

Hur skulle exemplet med blåmusslor ha sett ut med Kust som en slumpad faktor? Vi låtsas nu att vi plockat musslor från olika kustavsnitt men utan att vi haft något syfte att jämföra mellan några specifika områden. Syftet har bara varit och se hur mycket av variationen i längd som beror på att de växer på olika platser. Vi tar därför 2 stickprov på slumpmässigt valda kustavsnitt (egentligen för få i den här typen av undersökning). Vi ser att det är en slumpad faktor eftersom våra 2 nivåer bara är ett litet stickprov av alla tänkbara platser. Modellen ser ungefär likadan ut som ovan:
$L_{ij} = \mu + Plats_i + e_{ij}$

Vi antar nu också att vi får likadana stickprov av blåmusslor som ovan så att SS blir den samma. Det är nu skillnaden mellan fixerade och slumpade faktorer kommer in, och det gäller tolkningen. Faktorn Plats visar sig bidra till variationen i längd med 44 %. Men det vore meningslöst att tala om någon behandlingseffekt av Plats, eftersom det finns oändligt antal tänkbara platser som vi inte knyter någon särskild innebörd till. Olika platser betyder inget mer för oss än att de ligger på olika ställen. Två skäl finns ofta för att inkludera en slumpad faktor som i exemplet ovan:

1. Öka domänen av slutsatsen om hur musslors längd varierar
2. Bestämma den rumsliga variationen hos musslors längd

Regression

Tankegången kring statistiska modeller kan på ett tydligt sätt illustreras med en s.k. regressionsanalys. Ofta är man intresserad av att analysera relationen mellan flera faktorer. Ett exempel är en studie av populationstillväxten hos en ciliat i olika temperaturer. Vid några olika temperaturer skattar vi tillväxten, t.ex. r och vi får ett antal datapar.

Temperatur	Tillväxt (h^{-1})
2	0.24
4	0.35
6	0.43
8	0.5
10	0.72
12	1.1

Det verkar klart att det finns en relation mellan de två variablerna. Det är 2 saker vi kan vara intresserade av, troligen båda.

1. Beskriva det funktionella sambandet, d.v.s. kunna förutsäga tillväxten om vi vet temperaturen
2. Statistiskt testa H_0 om att det inte finns någon relation mellan temperatur och tillväxt

Båda dessa syften kan man nå med en regressionsanalys. Där passar man in en rät linje mellan datapunkterna enligt vissa regler. Det vanligaste är att använda sig av minsta kvadratmetoden för avstånden i y-led mellan punkterna och linjen. I korthet kan sägas att man sedan testar H_0 att lutningen är 0.

I vanlig regression gäller att vi testar ett funktionellt samband mellan två variabler, X och Y. Med en funktion mellan X och Y menas ju att om vi väljer ett X så är Y bestämd. X kallas därför för den oberoende variabeln och Y för den beroende. Även om en statistisk regression inte alltid handlar om orsak och verkan mellan X och Y i strikt mening, ska X vara något som kontrolleras av observatören. I exemplet ovan med tillväxt hos en ciliat i olika temperaturer, kontrollerar vi temperaturen. Vid ett antal utvalda temperaturer skattar vi tillväxthastigheten. Regressionen antar att temperaturen mäts utan fel. För varje temperatur skattar vi den beroende variabeln, tillväxthastigheten, ett antal gånger. Vi finner då att den blir något olika varje gång. Varje Y-värde är i själva verket ett stickprov ur en fördelning (Fig. 2).

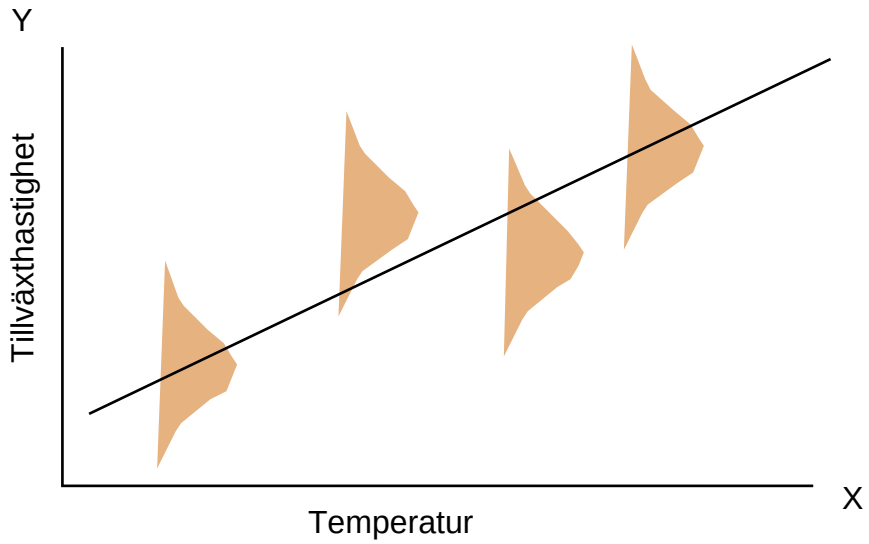

Fig. 2 Schematisk bild av spridningen av Y på olika nivåer av X.

Den linje vi nu anpassar är också en modell för att förklara variationen i de data vi har. Först kan vi formulera om medelvärdet av Y som en funktion av X.

$\mu = \alpha + \beta X$

där α är skärningspunkten på x-axeln och β är lutningskoefficienten ($\mu = \bar{Y}$ om $\beta = 0$). Nu kan vi formulera en modell som uttrycker varje enskilt Y-värde som summan av tre olika komponenter:

$Y_i = \alpha + \beta X + e_i$

Detta liknar mycket den föregående modellen med den skillnaden att vi förväntar oss ett linjärt beroende mellan X och Y. Det var inte fallet ovan med mussellängd och kust. Hur mycket av variationen i tillväxt förklarar då ett linjärt beroende av temperaturen, d.v.s. den räta linjen. Vi börjar med att se hur mycket oförklarad variation som finns runt medelvärdet ($\bar{Y}$) av tillväxten (Fig. 3a). Uttryckt som SS är den totala variationen mot det totala medelvärdet 0.484. Detta är SS för den totala variationen i Y-led. När vi sedan utökar vår modell till att omfatta beroendet av Temperatur, genom att använda de värden på α och β som passar bäst till data sjunker den oförklarade variationen (SS) till 0.055 (Fig. 3b). Regressionen, d.v.s. modellen som uttrycker Y som en funktion av X, förklarar i detta fall hela 88 % av den uppmätta variationen i Y ((0.484 - 0.055) / 0.484). Den variation som förklaras av den räta linjen (som specificeras av α och β) kan sedan fås genom subtraktion av denna oförklarade variation från den totala variationen (Fig. 3c).

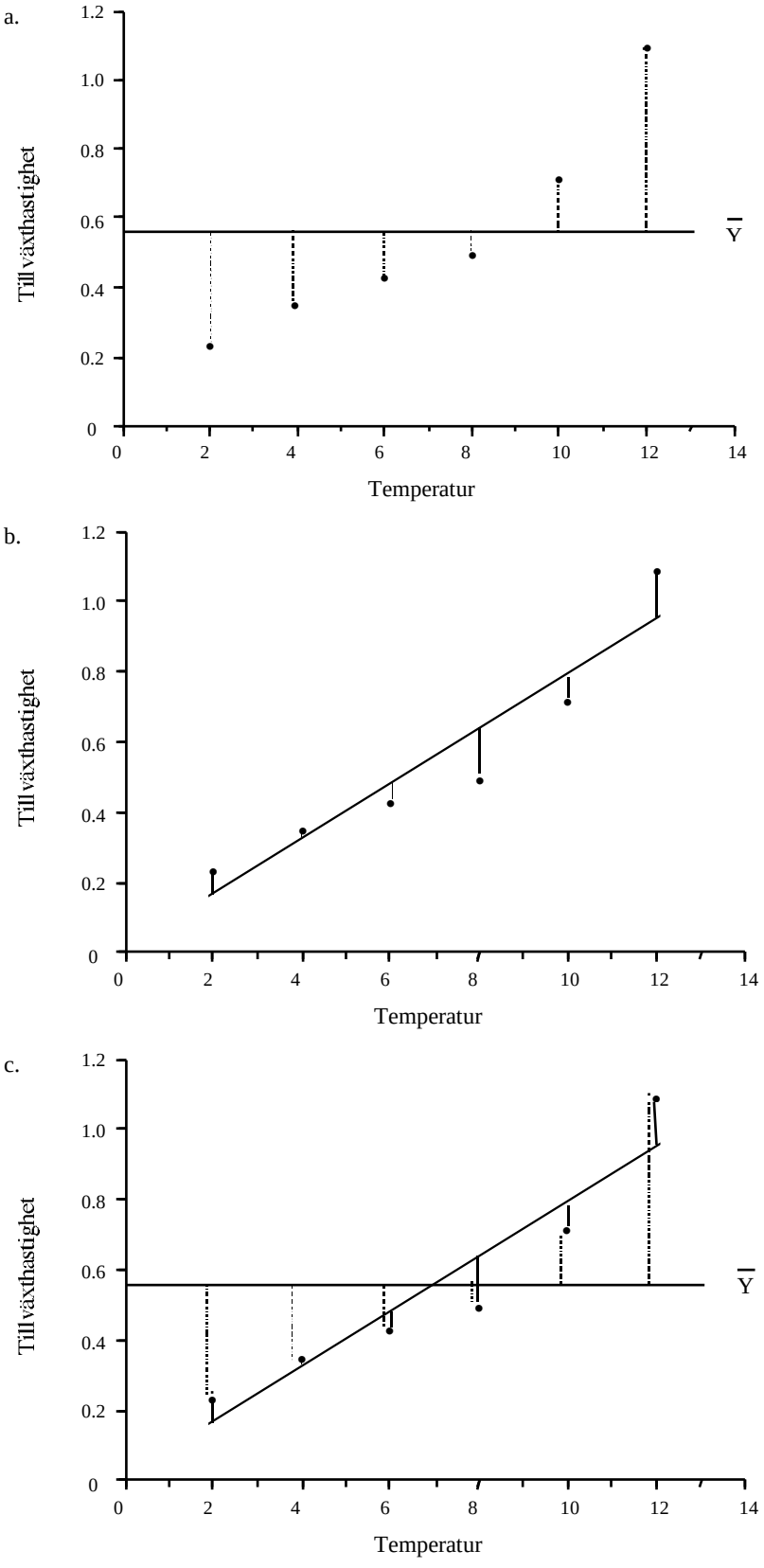

Fig. 3 Illustration av innebörden och beräkning av SS för regression. Se test för detaljer.

Man kan också nämna att regression kan utföras även med funktioner som inte är linjära. Ett exempel på sådana funktioner kan var polynom där varje enskilt Y-värde kan beskrivas som summan av fyra olika komponenter:

$Y_i = \alpha + \beta_1 X + \beta_2 X^2 + ei$

där α, β_1 och β_2 är skattade konstanter. Dessa typer av regressioner kommer dock inte att behandlas närmare här.

Hypotestestning

Vår regressionslinje förklarar visserligen 88 % av variationen i Y, d.v.s. det tycks som om det finns ett klart samband mellan variationen i X och Y. Men igen måste vi fråga oss om hur sannolikt detta utfall är. Eftersom vi använder stickprov finns en viss sannolikhet att det observerade mönstret har orsakats av slumpmässiga faktorer under experimentets gång eller vid provtagningen. Vi vill veta hur stor sannolikhet det är att vi drar ett stickprov av Y-värden som ger en regressionslinje, som vi har fått, om det inte finns något samband mellan X och Y i den verkliga populationen. Detta kan formuleras som att vi vill testa nollhypotesen att $\beta = 0$. Om nollhypotesen är sann är lutningen 0 och linjen går genom $\bar{Y}$ (Fig. 4). I praktiken kan detta innebära att man testar om SS för linjen är lika stor som för $\bar{Y}$. Vi ska strax se hur denna nollhypotes kan testas, men först en liten utvikning.

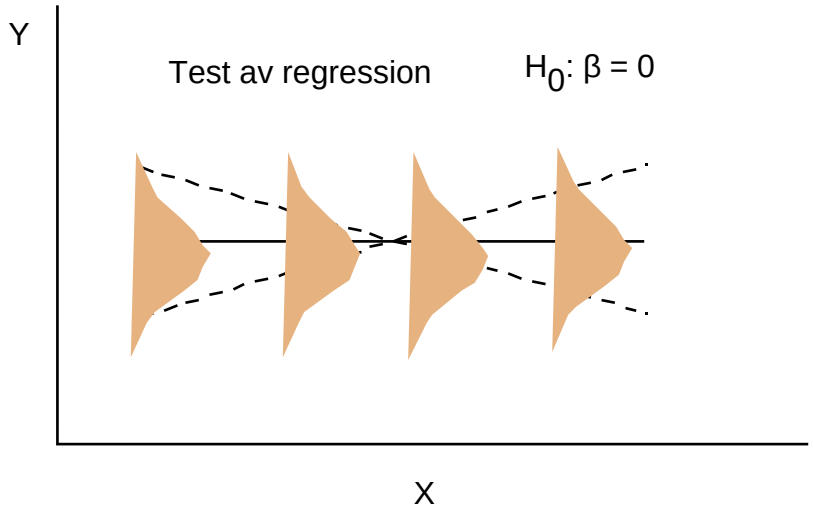

Fig. 4. Schematisk bild av spridningen kring medelvärdet om H_0 är sann.

F-kvot och variansanalys

Tidigare har vi haft 2 stickprov och vi har frågat oss med vilken sannolikhet de skattade medelvärdena kommer från samma population. Vi kan ställa oss samma fråga för 2 skattade varianser. Anta att vi har en känd normalfördelning och tar 2 stickprov. Om de är från samma fördelning borde vi förvänta oss att kvoten mellan de skattade varianserna är 1. På grund av icke-representativa stickprov kommer vi naturligtvis att få flera kvoter

som avviker från 1. Men hur stor avvikelse kan vi förvänta oss? Det är möjligt att räkna ut ett 95 % CI för kvoten. Om kvoten ligger utanför CI anser vi det vara osannolikt att varianserna kommer från samma fördelning och vi förkastar nollhypotesen.

Precis som i fallet med statistikan t vars fördelining är känd om det inte finns någon skillnad i medelvärden, är det också så att fördelningen av kvoten mellan två varianser är känd om det inte finns någon skillnad mellan varianserna. Denna statistika (kvoten mellan två varianser) kallas F (efter Ronald Fisher som utvecklade den) och har en fördelning som kallas F-fördelning. Notera nu att varje varians har sina df, så att en viss F-fördelning definieras av två df, en i täljaren och en i nämnaren.

$$\mathbf{F}_{v1,\,v2} = \frac{s_1^2(df = v1)}{s_2^2(df = v2)}$$

Vi är nu redo för ett stort steg. Vi skall nämligen göra en variansanalys. Om vi går tillbaka till vår regressionsanalys kan vi ställa upp följande tabell:

Faktor	d.f.	SS	MS	MS skattar	F
Linjen (Temp.)	1	0.429	0.429	$\sigma^2_e + \beta^2\Sigma x^2$	30.6
Ej förklarad	4	0.055	0.014	σ^2_e	
Totalt	5	0.484			

För att standardisera och kunna jämföra variationskomponenterna delas SS med d.f. och vi får det som kallas mean square, MS. MS är ett mått på variation som är oberoende av antalet mätningar. Man kan därför konstruera tabeller som är mycket mer allmängiltiga än om vi hade försökt skapa en statistika som baserades på SS direkt.

Om lutningen är noll förklarar regressionslinjen inte något ytterligare. Frågan är nu vad sannolikheten är att få den MS som i tabellen ovan om det inte finns något samband mellan X och Y, d.v.s. om nollhypotesen är sann. Om det faktiskt inte finns något samband mellan X och Y, kommer MS för linjen att skatta samman varians som MS för den ej förklarade variationen, d.v.s. σ^2_e. Nollhypotesen att $\beta = 0$ innebär att F-kvoten mellan de båda MS ska i genomsnitt vara 1. Om vi har ett bidrag från regressionslinjen ökar täljaren och vi får ett F som växer. Hur stort ska F vara för att vi ska dra slutsatsen att täljaren innehåller ett signifikant bidrag från regressionslinjen? Det kan vi se i en tabell över fördelningen av F-kvoten med 1 df i täljaren och 4 df i nämnaren. För $\alpha = 0.05$ är det kritiska F = 7.7. Vi har klart överstigit detta och vi förkastar H_0 och drar slutsatsen att det finns ett signifikant samband mellan X och Y. Nu har vi faktiskt gjort vår första variansanalys!

Modell I och Modell II regression

Genomgången av regression har hittills bara behandlat det fall när vi kontrollerar X och som är mätt utan fel. Denna typ av regression går ofta under namnet Model I. Inom biologin är det ofta så att även X är en skattad variabel. Som ett exempel kan vi tänka oss en undersökning av samvariationen mellan en predator och dess byte. Oftast används regression med en fixerad variabel här felaktigt. Det finns faktiskt andra regressionsmodeller som kan ta hänsyn till variation i både x och y-led, s.k. Modell II regression. Det finns också situationer när man inte har ett funktionellt samband mellan två variabler, men ändå vill anpassa en regression för att kunna förutsäga den ena från den andra. Även här kan Modell II användas. Exempelvis samvariation mellan morfologiska karaktärer. Ett problem med Modell II är att det inte finns någon vedertagen metod för att statistiskt testa regressionen. Enligt vissa statistiker kan man använda Modell I även för skattade X om man har många frihetsgrader. Det finns också icke-parametriska metoder t.ex. Spearmans rank korrelation.

Avsnittet om regression kan avslutas med att beröra ett snarlikt begrepp, korrelation. Dessa två metoder för att skatta samvariation leder ofta till förvirring och missuppfattningar. Korrelation beskriver samvariationen, hur stor del av variationen som beror på association. Är samvariationen mellan X och Y perfekt blir korrelationen 1, är den perfekt men omvänd blir korrelationen -1, och finns inget samband blir den 0. Korrelation kan statistiskt testas endast om man antar att stickprovet kommer ur en bivariat normalfördelning, vilket sällan uppfylls av biologiska data.