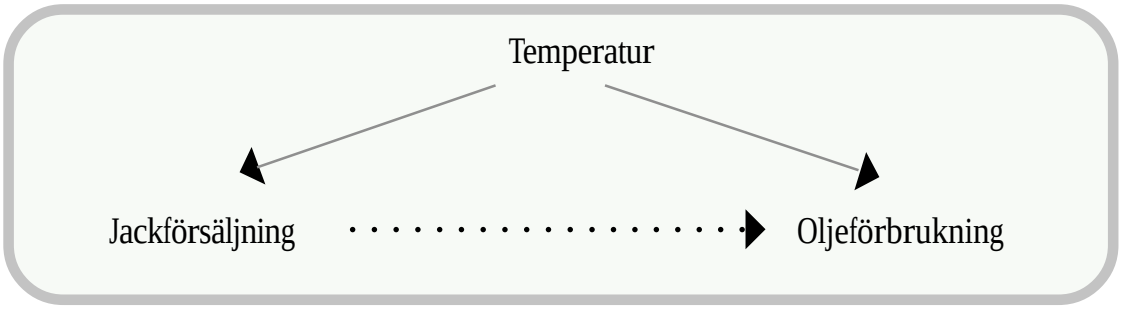
Multipel regression och Partiella korrelationer

Joakim Westerlund

Kom ihåg *bakomliggande variabelproblemet*:



Bakomliggande variabelproblemet kan, som tidigare nämnts, delvis angripas med hjälp av partiella korrelationer. På statistisk väg kan temperaturen hållas konstant. Om korrelationen mellan jackförsäljning och oljeförbrukning försvinner så vet vi att det var temperaturen som låg bakom det skenbara sambandet. Om korrelationen mellan jackförsäljning och oljeförbrukning kvarstår vet vi att temperaturen inte kan ligga bakom (däremot kan det finnas en oändlig mängd andra möjliga variabler som ligger bakom sambandet, mer om detta senare).

Ponera att vi slumpmässigt väljer ut 5 familjer och låter dessa under en månad spara alla kvitton gällande godis och mjölkinköp. För varje familj beräknar vi antalet kronor de spenderat på godis resp mjölk:

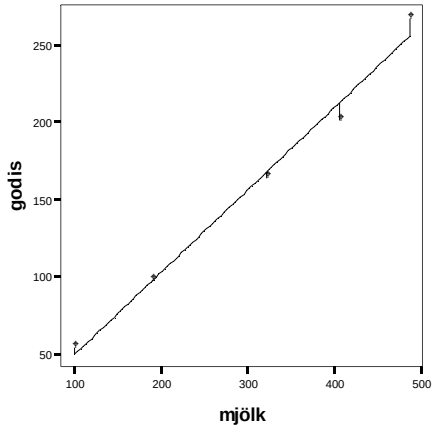
Mjölk	**Godis**
100	54
190	97
321	164
405	201
486	267

Det verkar finnas ett samband mellan hur mycket mjölk och hur mycket godis man köper. Vi beräknar r_{xy}:

Correlations

		GODIS
MJÖLK	Pearson Correlation	,995**
	Sig. (2-tailed)	,000
	N	5

**. Correlation is significant at the 0.01 level



Således ett starkt samband. (Behöver man mjölk för att skölja ned godiset? Passar godis som tilltugg till mjölk?)

Sedan noterar vi att familjerna har olika antal barn:

Mjölk	Godis	Antal barn
100	54	1
190	97	2
321	164	3
405	201	4
486	267	5

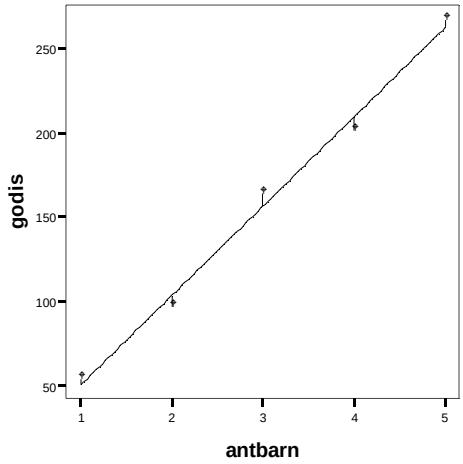
Undrar just om antal barn har nånting med godis och mjölkkonsumtion att göra?

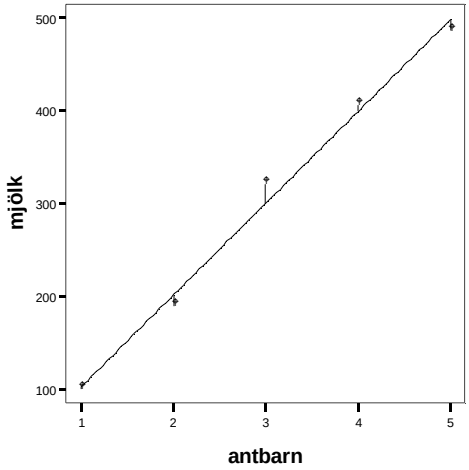
Correlations

		ANTBARN
MJÖLK	Pearson Correlation	,996**
	Sig. (2-tailed)	,000
	N	5
GODIS	Pearson Correlation	,996**
	Sig. (2-tailed)	,000
	N	5

**. Correlation is significant at the 0.01 level

Det hade det! Vi plottar dessa samband och sparar residualerna (Y – Y') som vi får vid prediktion från antbarn till mjölk resp. godis:



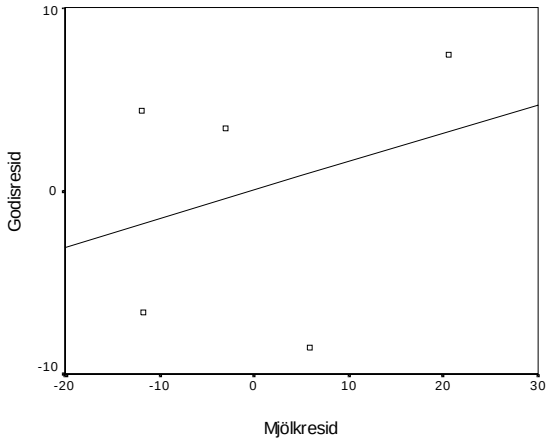


Mjölk	Godis	Antal barn	Mjölkresid	Godisresid
100	54	1	-3,00	3,40
190	97	2	-11,70	-6,60
321	164	3	20,60	7,40
405	201	4	5,90	-8,60
486	267	5	-11,80	4,40

Variationen i *mjölkresid* speglar variation i mjölkkonsumtion som inte har med antal barn att göra. Variationen i *godisresid* speglar variation i godiskonsumtion som inte har med antal barn att göra. Om vi korrelerar mjölkresid med godisresid får vi därför reda på hur starkt sambandet är mellan mjölk och godiskonsumtion sedan vi tagit bort allt som har med antal barn att göra:

Correlations

		Mjölkresid
Godisresid	Pearson Correlation	,300
	Sig. (2-tailed)	,623
	N	5



Som synes finns inte mycket kvar av sambandet! Slutsats: Antal barn var en bakomliggande variabel som gav upphov till ett skenbart samband mellan godis och mjölk konsumtion (Kanske finns det fler bakomliggande variabler som vi skulle kunna konstanthålla, t.ex. inkomst, och som skulle dra ner korrelationen ytterligare)

I stället för att hålla på och beräkna en massa residualer och korrelera dessa med varandra kan man använda en behändig formel för partiell korrelation. Korrelationen mellan variabel 1 och 2, med variabel 3 konstanthållen:

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

Multipel regression

Vid multipel regression används flera prediktorer (X_1, X_2, X_3,…) för att predicera en beroende variabel (Y). Ekvationen för prediktion av y från flera prediktorer ser ut så här:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n$$

Denna ekvation löses så att $\Sigma(Y - Y')^2$ blir så liten som möjligt. För att kunna utföra de faktiska beräkningarna utan hjälp av dator krävs kunskaper i matrisalgebra. Då sådana kunskaper inte kan förutsättas lämnar vi i det följande dessa beräkningar därhän och koncentrerar oss på hur multipel regression kan användas och tolkas.

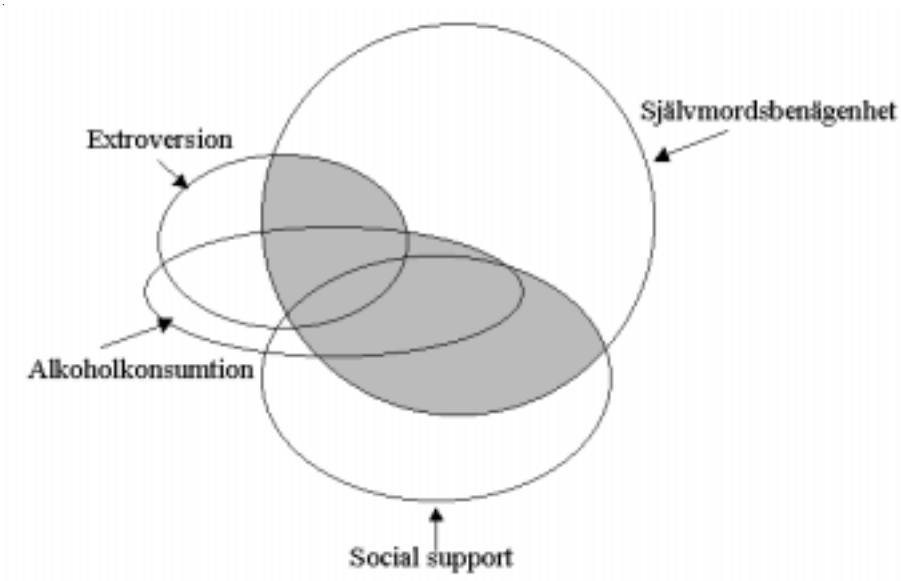
Prediktion

Ett typiskt användningsområde för multipel linjär regression är när man vill kunna göra så goda förutsägelser som möjligt vad gäller individers möjligheter att klara en utbildning, att bli bra piloter, att bli framgångsrika chefer osv. Resultat på ett antal test (X) används då som prediktorer av ett kriterium (Y) som vanligtvis föreligger först senare i tiden.

Förståelse

Ett annat användningsområde är när man vill förstå så mycket som möjligt av ett visst fenomen. Varför har t.ex. vissa människor en större benägenhet att begå självmord än andra? Detta skulle kunna undersökas genom att låta ett antal personer fylla i ett självmordsbenägenhetstest samt ett extroversionstest, ett test för mätning av grad av social support och ett test för mätning av alkoholvanor. Genom att se hur variationen i självmordsbenägenhet kan förklaras av variation i extroversion, social support och alkoholvanor kan man förhoppningsvis bättre förstå varför en del människor begår självmord och kanske t.o.m. använda denna kunskap för att minska antalet självmord.

Precis som man med den vanliga korrelationskoefficienten, r, kan beräkna sambandet mellan X och Y, kan man beräkna den multipla korrelationskoefficienten, R, mellan (X_1, X_2, X_3,…) å den ena sidan och Y å den andra. R^2 står för den del av variationen i Y som förklaras av X_1, X_2, X_3,… tillsammans. Exempelvis skulle R^2 kunna stå för den del av variationen i självmordsbenägenhet som kan förklaras av variation i extroversion, social support och alkoholkonsumtion:



Som synes är de oberoende variablerna sinsemellan högt korrelerade – något som kallas *multikolinearitet.* Notera att alkoholkonsumtionen inte nämnvärt ökar vår prediktionsförmåga om vi redan har med Extroversion och Social support som prediktorer, har vi bara med Extroversion ökar alkoholkonsumtionen däremot prediktionsförmågan avsevärt.

Exempel 1 på multipel regression med SPSS:

Några elever på psykologlinjen T1 gjorde en undersökning där de var intresserade av vilka faktorer som bestämmer om någonting är roligt. Deras försökspersoner fick bedöma ett antal sketcher på skalor 1-10. Förutom bedömd rolighet (Y) fick de skatta följande egenskaper (prediktorer):

(ofoer)*Oförutsägbarhet*: Hur pass mycket tycker du att handlingsförloppet i sketchen skiljer sig från det du förväntade dig i början av sketchen?
(identif) *Identifikation*: Hur mycket känner du igen dig i karaktärernas reaktioner?
(assoc) *Oväntade associationer*: I vilken grad tyckte du att sketchen innehöll oväntade associationer?
(tappakon) *Tappa kontrollen*: Till vilken grad tappade någon/några i sketchen kontrollen?
(normbr) *Normbrytare - upphållare*: I vilken grad tyckte du att sketchen följde ett mönster av att en uppehöll normen och en bröt den?
(utanf) *Utanförskap*: Till vilken grad skojar man med utanförskap i sketchen?
(igenk) *Vardagsigenkänning*: Hur mycket känner du igen dig i situationen från din egen vardag?
(story) *Story*: Hur mycket bygger humorn på att det berättas en historia?
(fara) *Fara*: I vilken grad uppfattar du att någon/några utsätts för fara?
(genans) *Genans*: Hur mycket bygger sketchen på att situationen är pinsam för någon/några?

Enskilda korrelationer mellan prediktorerna och bedömd rolighet:

Correlations

		ROLIG
OFOER	Pearson Correlation	,447**
	Sig. (2-tailed)	,000
	N	280
IDENTIF	Pearson Correlation	,393**
	Sig. (2-tailed)	,000
	N	279
ASSOC	Pearson Correlation	,538**
	Sig. (2-tailed)	,000
	N	280
TAPPAKON	Pearson Correlation	,051
	Sig. (2-tailed)	,395
	N	280
NORMBR	Pearson Correlation	,130*
	Sig. (2-tailed)	,030
	N	280
UTANF	Pearson Correlation	-,001
	Sig. (2-tailed)	,981
	N	280
IGENK	Pearson Correlation	,271**
	Sig. (2-tailed)	,000
	N	280
STORY	Pearson Correlation	,158**
	Sig. (2-tailed)	,008
	N	280
FARA	Pearson Correlation	-,069
	Sig. (2-tailed)	,247
	N	280
GENANS	Pearson Correlation	,029
	Sig. (2-tailed)	,627
	N	280

**. Correlation is significant at the 0.01 level

*. Correlation is significant at the 0.05 level (2-tailed).

Om vi i stället låter datorn (SPSS) beräkna linjär regression får vi följande resultat:

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	GENANS, IGENK, OFOER, STORY, FARA, NORMBR, UTANF, TAPPAKON, ASSOC, IDENTIF[a]	,	Enter

a. All requested variables entered.

b. Dependent Variable: ROLIG

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,673[a]	,452	,432	1,8599

a. Predictors: (Constant), GENANS, IGENK, OFOER, STORY, FARA, NORMBR, UTANF, TAPPAKON, ASSOC, IDENTIF

b. Dependent Variable: ROLIG

R2 är mer intressant än R. Med hjälp av våra prediktorer (ofoer, identif, … genans) kan vi alltså förklara 45.2% av variationen i bedömd rolighet. R2 är dock inget väntevärdesriktigt estimat av motsvarande populationsegenskap. Det är däremot ”Adjusted R Square”.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	765,831	10	76,583	22,139	,000[a]
	Residual	927,079	268	3,459		
	Total	1692,910	278			

a. Predictors: (Constant), GENANS, IGENK, OFOER, STORY, FARA, NORMBR, UTANF, TAPPAKON, ASSOC, IDENTIF

b. Dependent Variable: ROLIG

Till skillnad från vanliga r är icke det förväntade slumpvärdet på R = 0 utan p/(N – 1) där p = antalet prediktorer och N = antal individer.

Vi signifikansprövar den multipla korrelationen med ett F-test. Som synes blev det klart signifikant.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,611	,358		4,499	,000
	OFOER	,302	,055	,314	5,450	,000
	IDENTIF	,304	,064	,301	4,739	,000
	ASSOC	,278	,055	,301	5,065	,000
	TAPPAKON	-7,92E-02	,056	-,078	-1,423	,156
	NORMBR	9,019E-02	,041	,122	2,194	,029
	UTANF	-8,67E-02	,046	-,101	-1,887	,060
	IGENK	-2,89E-03	,060	-,003	-,048	,962
	STORY	5,450E-02	,047	,055	1,157	,248
	FARA	-,322	,099	-,156	-3,258	,001
	GENANS	5,884E-03	,054	,006	,110	,913

a. Dependent Variable: ROLIG

Kolumnen B under ”Unstandardized coefficients” visar vikterna för vår regressionsekvation. ”Constant” = interceptet, som vi kallat ”a” tidigare. Vi får alltså:

$$\hat{Y} = 1.611 + \\ + .303\times OFOER \\ + .304\times IDENTIF \\ + .278\times ASSOC \\ - .0792\times TAPPAKON \\ + .09019\times NORMBR \\ - .0867\times UTANF \\ - .00289\times IGENK \\ - .322\times FARA \\ + .005884\times GENANS$$

Hur ”viktiga” de olika prediktorerna är kan man dock förstå lättare genom att titta i kolumnen ”Standardized coefficients”. Dessa anger regressionsvikterna för prediktion av Z_Y, när alla prediktorer är z-transformerade. De standardiserade regressionsvikterna brukar betecknas ?. Låt oss titta på det första värdet 0.314. Detta säger oss att en ökning av bedömd oförutsägbarhet med en enhet (Z), *om alla andra prediktorer hålls konstanta*, kommer att leda till en ökning av predicerad rolighet med 0.314 enheter (Z).

Av avgörande betydelse är naturligtvis om prediktorerna ger signifikanta bidrag till förklarandet av variationen i Y. Detta kan vi inspektera i sista kolumnen.

OFOER, IDENTIF, ASSOC, NORMBR och FARA var alla signifikanta, UTANF närmade sig signifikans. Jämför detta med de enkla korrelationerna. För dessa var OFOER, IDENTIF, ASSOC, NORMBR, IGENK och STORY signifikanta. Om vi har med alla prediktorer samtidigt bidrar uppenbarligen inte IGENK och STORY till något nytt, antagligen på grund av att vad dessa variabler egentligen mäter till stor del redan ingår i de andra variablerna. (IGENK mäts nog av IDENTIF).

Exempel 2 på multipel regression med SPSS:

För 25 personer har vi tillgång till deras gymnasiebetyg, deras resultat på högskoleprovet och deras framgångar på högskolan (på en 8-gradig skala!). Ponera att om vi korrelerar alla variabler mot varandra så får vi:

Correlations

		BETYG	HÖGPROV	FRAMGÅNG
BETYG	Pearson Correlation	1	,755**	,675**
	Sig. (2-tailed)	,	,000	,000
	N	25	25	25
HÖGPROV	Pearson Correlation	,755**	1	,635**
	Sig. (2-tailed)	,000	,	,001
	N	25	25	25
FRAMGÅNG	Pearson Correlation	,675**	,635**	1
	Sig. (2-tailed)	,000	,001	,
	N	25	25	25

**. Correlation is significant at the 0.01 level (2-tailed).

Och om vi beräknar multipel regression med betyg och resultat på högskoleprovet som prediktorer och framgångar på högskolan som kriterium med SPSS så får vi följande resultat:

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	BETYG, HÖGPROV[a]	,	Enter

a. All requested variables entered.

b. Dependent Variable: FRAMGÅNG

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,702[a]	,492	,446	1,387

a. Predictors: (Constant), BETYG, HÖGPROV

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	41,050	2	20,525	10,673	,001[a]
	Residual	42,310	22	1,923		
	Total	83,360	24			

a. Predictors: (Constant), BETYG, HÖGPROV

b. Dependent Variable: FRAMGÅNG

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	,487	,843		,577	,570			
	HÖGPROV	,422	,335	,291	1,259	,221	,635	,259	,191
	BETYG	,679	,345	,455	1,966	,062	,675	,387	,299

a. Dependent Variable: FRAMGÅNG

Här bad jag SPSS om de partiella (kolumnen Partial) och de semipartiella (kolumnen Part) korrelationerna mellan prediktorerna och kriteriet.

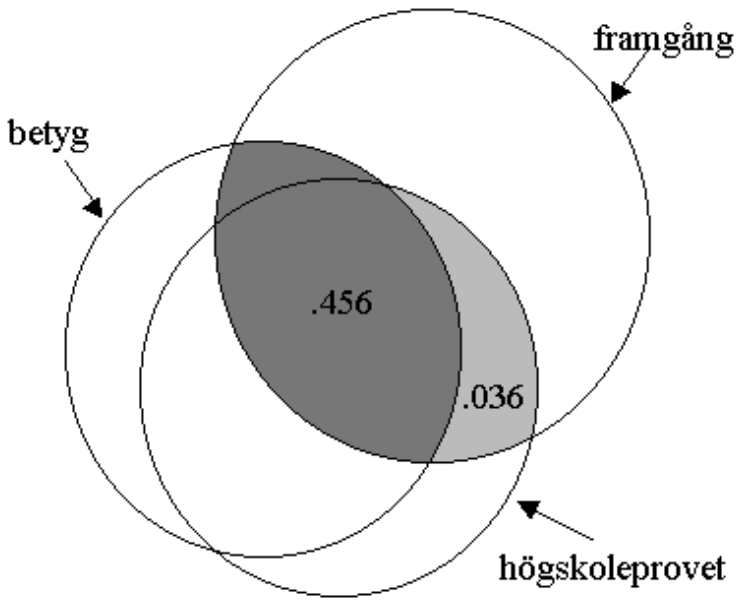
Den partiella korrelationen mellan högskoleprovet och framgång (0.259) innebär korrelationen mellan högskoleprovet och framgång sedan effekten av betyg tagits bort (hållits konstant) från både högskoleprovet och framgång.

Den semipartiella korrelationen mellan högskoleprovet och framgång (0.191) innebär däremot korrelationen mellan högskoleprovet och framgång sedan effekten av betyg tagits bort (hållits konstant) från bara högskoleprovet.

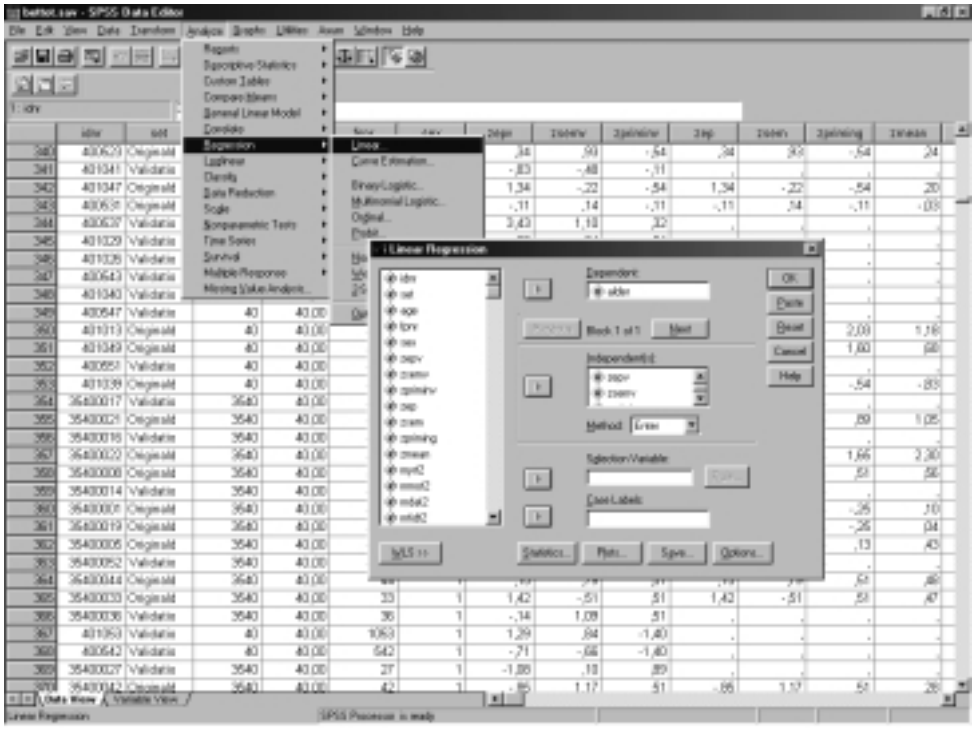
Den partiella korrelationen mellan betyg och framgång (0.387) innebär korrelationen mellan betyg och framgång sedan effekten av högskoleprovet tagits bort (hållits konstant) från både betyg och framgång.

Den semipartiella korrelationen mellan betyg och framgång (0.299) innebär däremot korrelationen mellan betyg och framgång sedan effekten av högskoleprovet tagits bort (hållits konstant) från bara betyg.

Den kvadrerade multipla korrelationskoefficienten ($.702^2$ = 0.492) kan ses som summan av den kvadrerade korrelationen mellan betyg och framgång (0.675^2 = 0.456) och den kvadrerade semipartiella korrelationen mellan högskoleprovet och framgång (0.191^2 = 0.036). Se figuren.



Exempel 3 på multipel regression med SPSS:



Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	ZPRIMINV, ZSEMV, ZEPV[a]	,	Enter

a. All requested variables entered.

b. Dependent Variable: ALDER

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,669[a]	,448	,447	10,5001

a. Predictors: (Constant), ZPRIMINV, ZSEMV, ZEPV

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	240381,8	3	80127,251	726,760	,000[a]
	Residual	296359,3	2688	110,253		
	Total	536741,0	2691			

a. Predictors: (Constant), ZPRIMINV, ZSEMV, ZEPV

b. Dependent Variable: ALDER

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	51,807	,270		191,625	,000
	ZEPV	-7,321	,201	-,717	-36,389	,000
	ZSEMV	,968	,227	,083	4,259	,000
	ZPRIMINV	-,504	,223	-,033	-2,264	,024

a. Dependent Variable: ALDER