Inlämningsuppgift 2, del 2A förutspå medeltemperatur

Modell för att förutspå medeltemperaturen på en plats på jorden.

Följande ekvation är en polynommodell för att förutspå medeltemperaturen för en plats på jorden utifrån dess Altitud i meter och Latitud utryckt i decimalform. Modellen togs fram med tilldelad data som innehöll variablerna temperatur, stad, land, kontinent, altitud, avstånd till havet, latitud, riktning från ekvatorn, longitud och riktning från nollmeridianen.

$$Medeltemperatur\ i\ Celcius = 27.6 - 0.00429 * Altitud + 0.0126 * Latitud - 0.00731 * Latitud^2$$

Denna slutgiltiga modell som togs fram ur datan, som även kan ses i tabell 2 (Kol, 3), använder *Altitud* och *Latitud* för att predikera *Medeltemperaturen i Celcius* enligt dem värdena. Detta betyder alltså att, enligt vår modell, minskar medeltemperaturen med 0.00429 grader för varje meter över havet platsen ligger. Vidare, koefficienten framför $Latitud^2$ tillsammans med koefficienten framför $Latitud$ mäter effekten av avståndet från ekvatorn på medeltemperaturen. Latituds påverkan på medeltemperaturen ökar desto högre latituds värde är, dvs . Detta kan ses rent grafiskt i figur 2, samt numeriskt i tabell 1 som visar tre olika lutningar för tre olika latituder.

Lutning vid Latitud 20	Lutning vid Latitud 40	Lutning vid Latitud 60
-0,2798	-0,5722	-0,8646

Tabell 1

	(1)	(2)	(3)	(4)	(5)
VARIABLER	Linjär med Altitud	Polynom utan Altitud	Polynom med Altitud	Polynom med Altitud under 2000 meter	Polynom med Altitud under 1500 meter
Altitud	-0.00398**		-0.00429**	-0.00395**	-0.00284**
	(0.000578)		(0.000360)	(0.000651)	(0.000918)
Latitud	-0.411**	-0.0925	0.0126	0.00458	-0.0717
	(0.0265)	(0.0625)	(0.0829)	(0.0899)	(0.0922)
Poly_Lat		-0.00531**	-0.00731**	-0.00721**	-0.00612**
		(0.00102)	(0.00152)	(0.00161)	(0.00165)
Konstant	31.72**	26.57**	27.60**	27.66**	28.49**
	(1.052)	(0.954)	(0.794)	(0.879)	(0.905)
Observationer	50	143	50	47	44
R-squared	0.845	0.841	0.907	0.908	0.913
Robusta standardfel inom paranteserna ** p<0.01, * p<0.05					

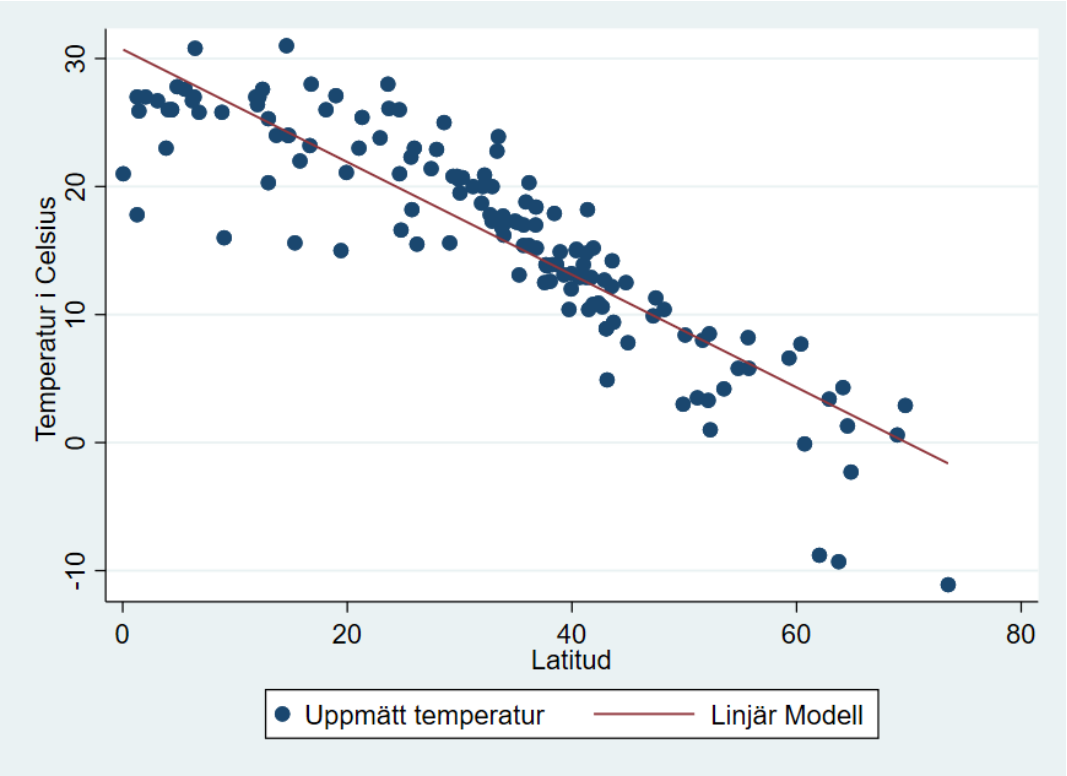
Tabell 2

Hur modellen togs fram.

Med den tillgängliga datan ansågs det, rent teoretiskt, att endast variablerna temperatur, latitud och altitud var relevanta för vår analys. Stad, land, kontinent, longitud, öst, väst och avstånd till havet

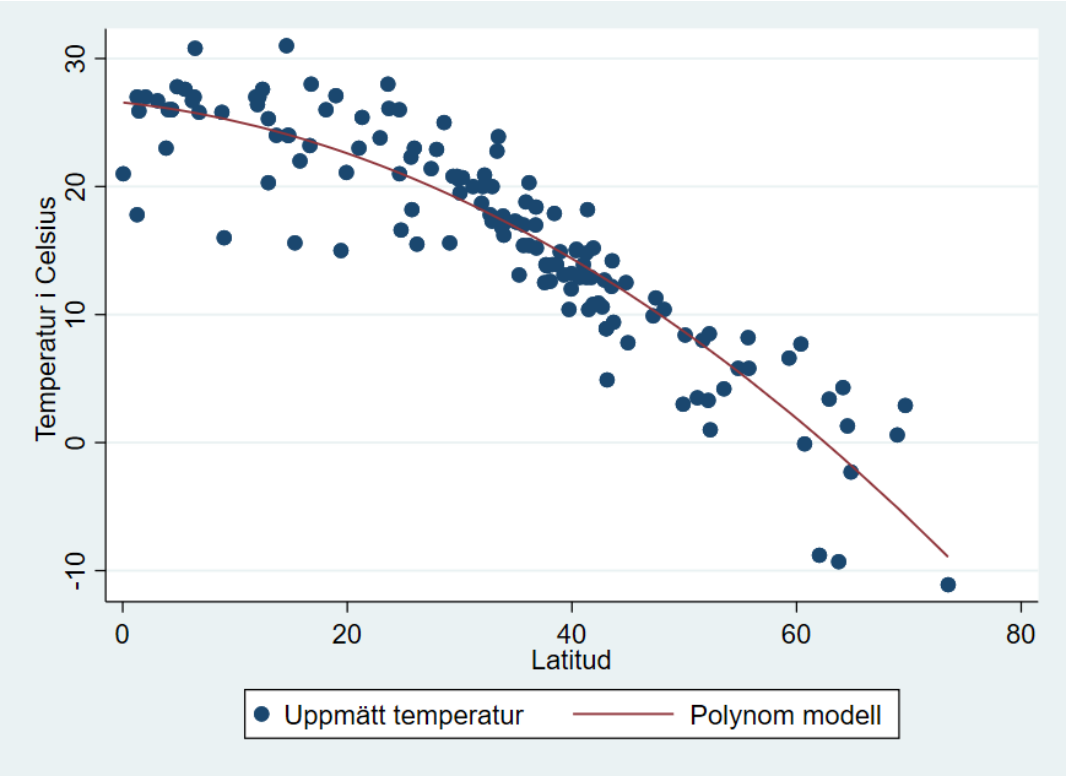
sågs inte som att de tillförde något till analysen och är därför inte med i den slutgiltiga modellen (Tabell 2, kol 3). Detta eftersom medeltemperaturen för en plats på jorden i teorin varierar primärt utifrån hur långt platsen ligger ifrån ekvatorn och platsens höjd över havet. Då den informationen finns i variablerna latitud samt altitud, var inte longitud, nord, öst, stad, land och kontinent relevanta till analysen. Vidare är det alltså inte heller av betydelse vilken specifik riktning ifrån ekvatorn en plats ligger, endast hur långt ifrån ekvatorn platsen ligger.

Slutligen, efter bortsorterandet av de tidigare nämnda irrelevanta variablerna, undersöktes förhållandet mellan temperatur och latitud, och förhållandet mellan temperatur och altitud för sig.



Figur 1, linjär modell

I figur 1 ovan syns förhållandet mellan temperatur och latitud och en linjär prediktion (tabell 3, kol 1). En ser tydligt att förhållandet inte tycks vara helt linjärt utan att medeltemperaturen verkar minska snabbare desto längre ifrån ekvatorn en plats ligger. Av denna anledning testades en polynommodel (tabell 2, kol 2) med kvadrerad latitud. Denna syns nedan och tycks passa datan mycket bättre (Figur 2).

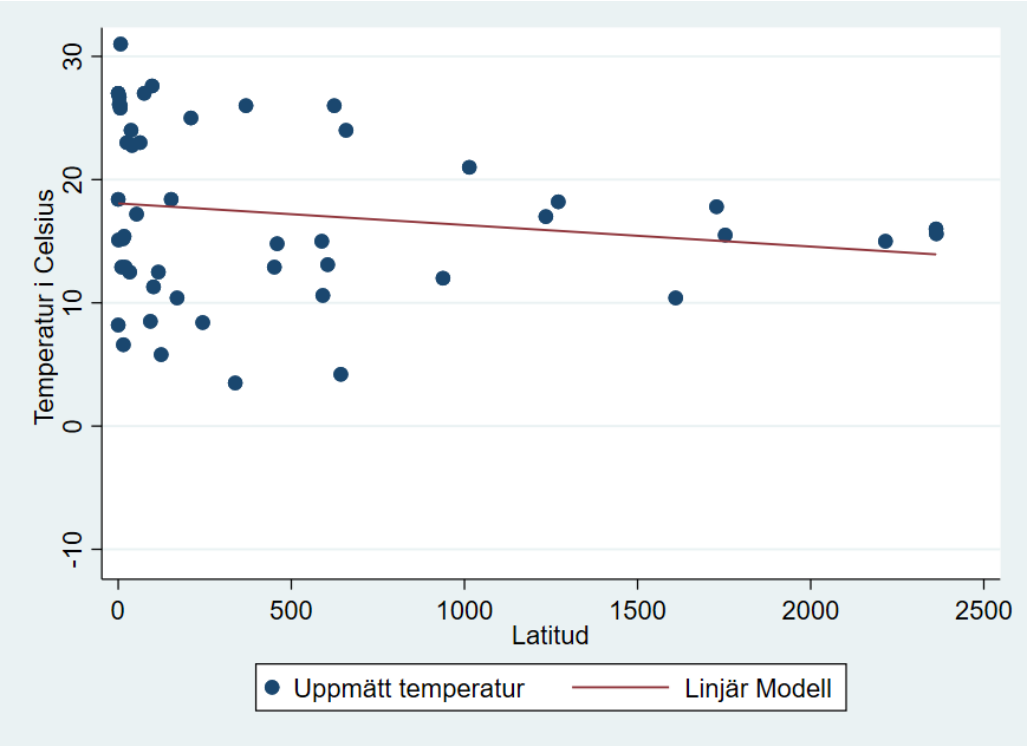


Figur 2, icke-linjär modell

Efter latitud tittades det på förhållandet mellan temperatur och variabeln altitud (figur 3) tillsammans med en linjär prediktion (tabell 3, kol 2). När detta gjordes märktes det att spridningen i variabeln altitud var ojämn och att de allra flesta observationerna låg i vänstra halvan vilket gör att de få observationer till höger potentiellt får orimligt stor påverkan på regressionslinjen. Samtidigt är denna fördelning representativ eftersom datan kommer från städer och de flesta städer ligger inte särskilt högt över havet. Vidare, att temperaturen minskar när altituden ökar är allmänt känt, men det är svårare att dra någon slutsats om huruvida denna svaga lutning som syns är rimlig eller inte. Det finns alltså en chans att den svaga lutningen skulle kunna vara ett resultat av slumpen.

	(1)	(2)
	Linjär utan Altitud	Linjär utan Latitud
VARIABLER		
Latitud	-0.440**	
	(0.0250)	
Altitud		-0.00175
		(0.000908)
Konstant	30.72**	18.07**
	(0.891)	(1.285)
Observationer	143	51
R-squared	0.791	0.027
Robusta standardfel inom paranteserna		
** p<0.01, * p<0.05		

Tabell 3

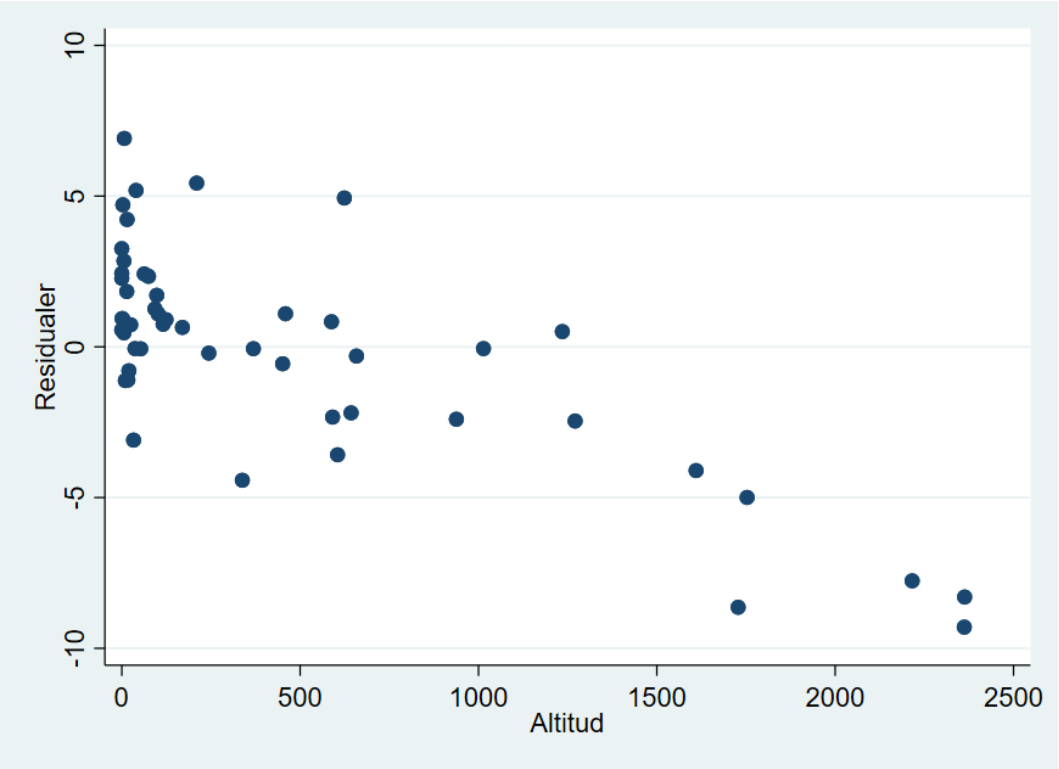


Figur 3

I figur 4 visas residualerna från regressionen $Temperatur = Latitud * x_1 + Latitud^2 * x_1$ i ett spridningsdiagram tillsammans med altitud. I diagrammet syns ett mönster som visar att residualerna blir mer negativa med högre altitud vilket innebär att temperaturen minskar med mer desto högre altitud än vad som kan förklaras av endast latitud.

Detta samband är som förväntat då det är allmänt känt att på hög altitud är det kallare än på låg altitud. Å andra sidan, som beskrevs tidigare i rapporten, med tanke på hur få observationer för platser på hög altitud som fanns med i datan kan slumpen haft orimligt stor påverkan. Därför testades även altituds påverkan på temperatur med endast de observationerna med altitud under 2000 meter (tabell 2, kol 4) och sedan endast med de under 1500 meter (tabell 2, kol 5). I båda fallen blev altituds påverkan på temperatur likt tidigare både negativ och signifikant men effekten av altitud minskade, speciellt utan de observationerna från över 1500 meter.

Ett annat problem med variabeln altitud var att den endast hade värden angivna för ca en tredjedel av observationerna. Detta betydde att när den inkluderades förlorades ca två tredjedelar av alla möjliga observationer. Därför behövdes beslutet tas huruvida en viktig variabel eller fler observationer var mer betydelsefullt för analysen. Efter en avvägning beslutades det att variabeln altitud med alla dess observationer skulle inkluderas i den slutgiltiga modellen eftersom målet var att skapa en modell för att förutspå medeltemperatur varsomhelst på jorden. Vidare, då altitud varierar kraftigt mellan olika platser på jorden och detta i sin tur påverkar medeltemperaturen på dessa platser, framstod den modell med altitud som den mest optimala modellen även om antalet använda observationer blev mindre.



Figur 4

Utvärdering av modellen

Slutligen sågs det som att Polynommodellen med altitud inkluderad är den bästa modellen (tabell 2, kol 3) eftersom den hade bättre passform än de linjära modellerna. Det är däremot viktigt att notera att den även har svagheten att det blev mycket färre observationer kvar att basera modellen på när altitud inkluderades. Med det i åtanke kanske polynommodellen utan altitud inkluderad (tabell 2, kol 2) är bättre för att förutspå medeltemperatur för en plats som inte ligger högt över havet. Huruvida de två andra polynommodellerna med altitud där de högre altitudobservationerna exkluderades (tabell 2 kol 4,5) är bättre än den där det inte gjordes är okänt. Detta eftersom det inte är känt om de högre altitudvärdena är representativa eller ett resultat av slumpen.

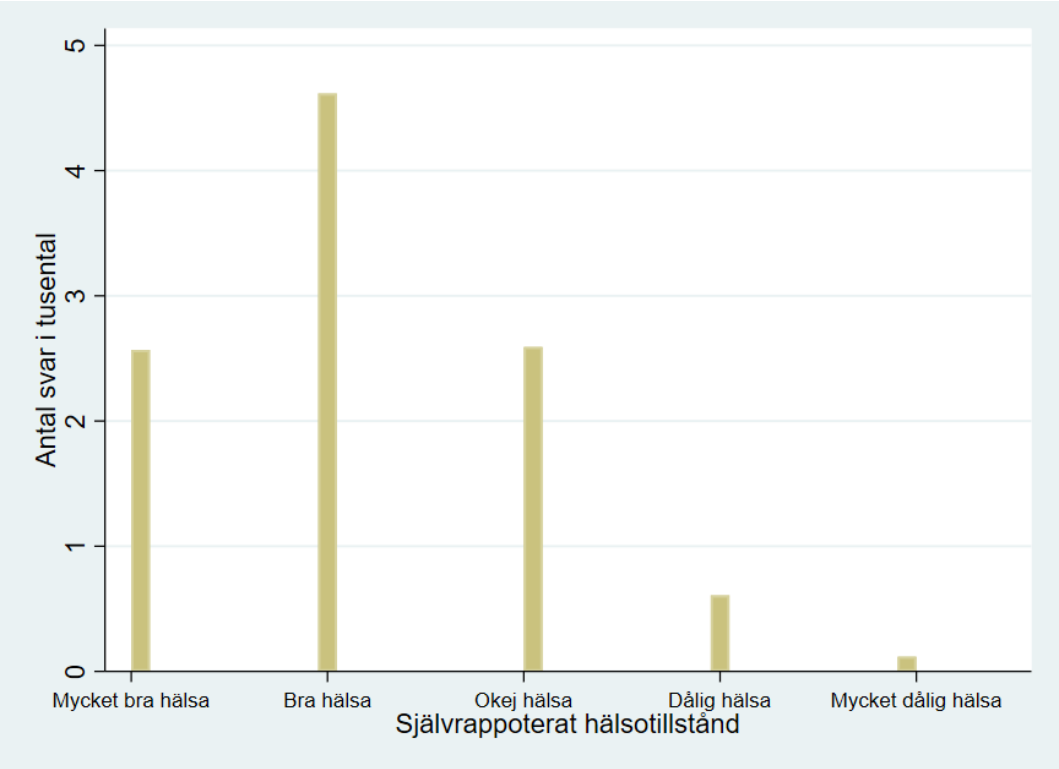
Oavsett skulle fler observationer vara önskvärt i båda fallen och en användare av modellen bör inte använda någon av modellerna till att göra något annat en grov uppskattning.

Inlämningsuppgift 2B

Följande rapport analyserar sambandet mellan självrapporterad hälsa, ålder och BNP/C baserat på ett urval av data från European Social Survey 2016. Målet är att ta fram en rimlig modell som beskriver sannolikheten för att ha god hälsa utifrån tidigare nämnda variabler. För att uppnå detta har vi analyserat data från ett antal länder som skiljer sig mellan 11300 BNP/C (Polen) och 68200 BNP/C (Norge). Vidare, är all data insamlad från Europa och hälsan har mätts genom självrapportering av hälsa uppdelat i kategorierna: Mycket god hälsa, god hälsa, okej hälsa, dålig hälsa och mycket dålig hälsa.

Eftersom den beroende variabeln hälsa är kategorisk behövdes det antingen göras en modell för varje kategori eller så behövde kategorierna grupperas. Med tanke på att sambanden blir krångligare att analysera och att det inte tillför särskilt mycket mer att rapportera en modell för varje hälsokategori, så togs beslutet att gruppera kategorierna i två nya kategorier. Dessa två kategorierna var 'God hälsa', som inkluderar alla kategorier över samt 'okej hälsa', och 'dålig hälsa' vilken inkluderar de kategorier under 'okej hälsa'.

Denna indelning framstod logisk då det känns som att steget mellan 'okej hälsa' och 'dålig hälsa' är större än steget mellan 'okej hälsa' och 'god hälsa' rent intuitivt. Dessutom, om man kollar på histogrammet (graf 1) så är det mycket vanligare att rapportera åtminstone 'okej hälsa' än 'dålig hälsa'. Detta betyder att observationerna från 'okej hälsa' skulle ha en väldigt stor inverkan på kategorin 'dålig hälsa' om indelningen var åt andra hållet.



Graf 1, Histogram som visar fördelningen av självrapporterad hälsa

Som nämndes i ovanstående paragraf är vår beroende variabel "hälsa" kategorisk. Detta gör analysen av en linjär modell komplicerad då även om dess tolkningsbarhet är överlägsen så är inte dess passningsförmågan det. Vidare är inte det verkliga sambandet mellan ålder, BNP/C och hälsa antagligen linjärt heller, detta betyder att det inte kanske inte så relevant att se i siffror hur mycket chansen att ha god hälsa minskar linjärt beroende på ålder och BNP/C. Istället bör det ligga en fokus

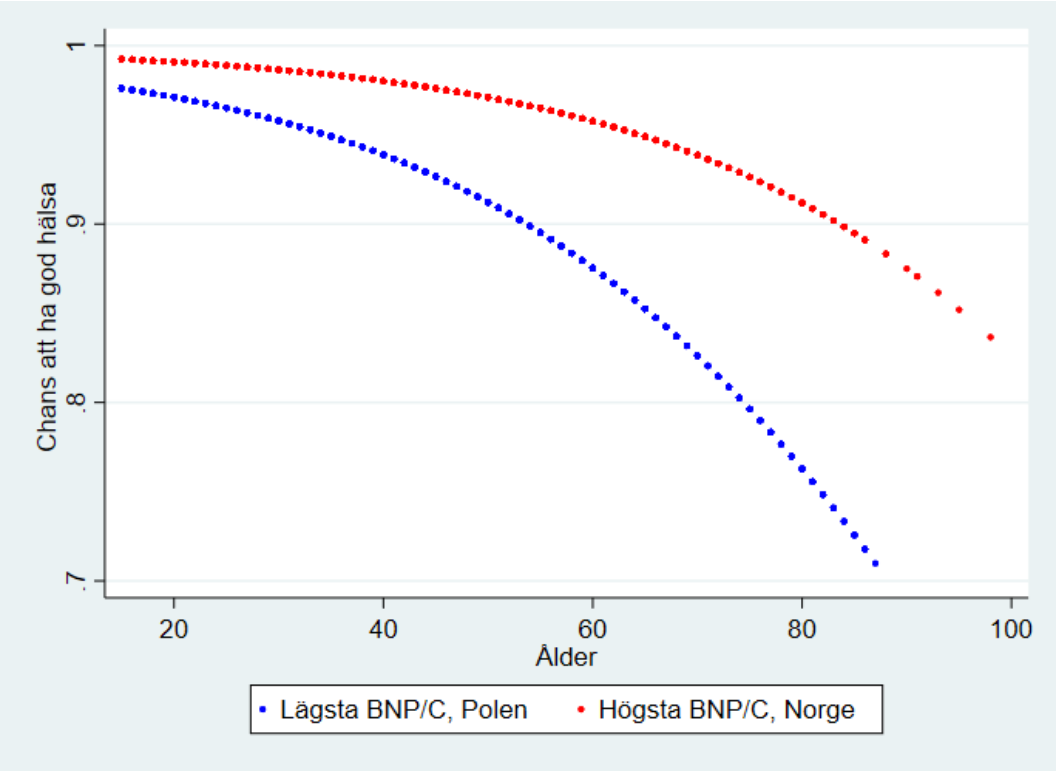
på om sambandet är negativt eller positivt. Det mest rimliga blir då att istället använda en logit modell med vår hälsovariabel och våra oberoende variabler ålder och BNP/C och att istället primärt rapportera sambanden rent grafiskt. Vidare rapporterar vi senare i rapporten även ett antal punkter för att förmedla hur stor chans det är att vara vid god hälsa för en specifik ålder i högt, lågt och medel BNP/C.

Ett annat problem som framkommer i vår analys är att vi jämför variabler som varierar på både individ nivå (hälsa, ålder) samt land nivå (BNP/C). Detta orsakar en inomgruppskorrelation i vår data då våra observationer från samma land har betydligt mer gemensamt än bara samma BNP/C. Rent praktiskt innebär detta att vi egentligen bara har 15 observationer på BNP/C nivån och inte 16,876 (Tabell 1). Detta leder till överdriven säkerhet av skattningarna i vår analys. Därför valdes det att använda klustrade standardfel för att inte våra p-värden skulle bli för låga, och våra standardfel för små.

VARIABLER	Logit modell för god hälsa med klustrade standardfel
Ålder	-0.0391**
	(0.00457)
BNP/C	2.05e-05**
	(6.68e-06)
Konstant	4.062**
	(0.274)
Observationer	16,876
Robusta standardfel inom paranteserna	
** p<0.01, * p<0.05	

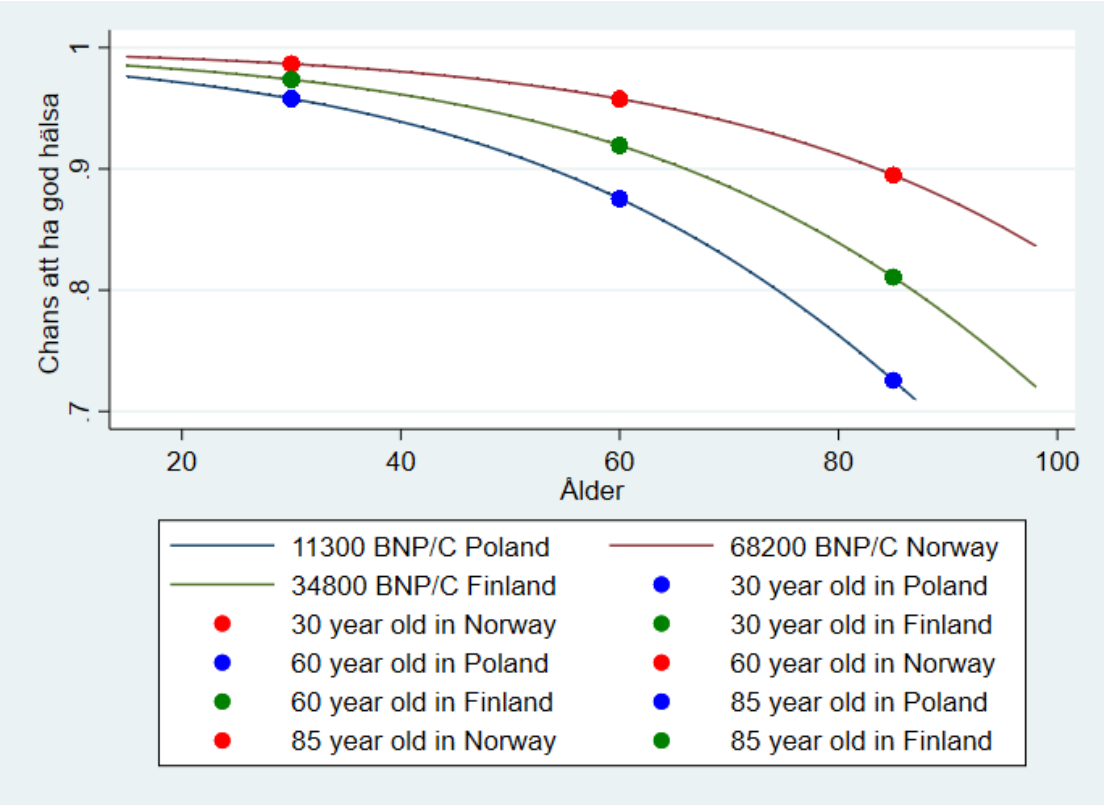
Tabell 2, regression för Hälsa, ålder och BNP/C

Som man kan se i tabell 1, även om siffervärdena på koefficienterna i en logit modell inte lätt kan tolkas, finns det ett tydligt negativt samband mellan hög ålder och chans att vara vid god hälsa och ett tydligt positivt samband mellan ett högt BNP/C och chansen att vara vid god hälsa. Detta kan även illustreras grafiskt som i graf 2 där prediktionerna från vår logit modell (Tabell 1) för att ha god hälsa utifrån ålder för två olika BNP/C nivåer visas. En kan här tydligt se att chansen att vara vid god hälsa inte skiljer sig avsevärt mycket mellan unga människor i länder med olika BNP/C nivåer. Tittar man däremot på kurvorna märker en att skillnaderna i lutning mellan Polen och Norge blir betydligt större ju äldre en person blir, vilket betyder att chansen att vara vid god hälsa vid slutet av kurvan är mycket mindre för invånare i Polen (Graph 2). Därmed blir antalet vid god hälsa lägre i de högre åldrarna för länder med lågt BNP/C.



Graf 2, graf som visar alla observationer från Polen och Norge

I graf 3 har prediktionerna för åldrarna 30, 60 och 85 från vår modell valts ut och markerats på kurvorna för den högsta, mellersta och lägsta BNP/C nivån. Chansen att ha god hälsa vid dessa tre åldrar utifrån BNP/C nivå är listad i tabell 2 och kurvanslutning, dvs hur mycket chansen minskar när en person blir ett år äldre än den listade åldern, finns i tabell 3. Dessa tabeller bekräftar vad som sågs redan i graf 2, alltså att sannolikheten att vara vid god hälsa utifrån BNP/C skiljer sig inte mycket för unga men är desto större för äldre. Vidare minskar chansen allt snabbare med åldern för människor i länder med lågt BNP/C.



Graf 5, graf som visar Norge, Finland och Polen samt 9 punkter för 3 olika åldrar i varje land som visar chansen att ha god hälsa.

	BNP/capita		
	11300 (Poland)	34800 (Finland)	68200 (Norge)
Chans att ha god hälsa vid 30 års ålder	95,78%	97,35%	98,65%
Chans att ha god hälsa vid 60 års ålder	87,54%	91,92%	95,76%
Chans att ha god hälsa vid 85 års ålder	72,60%	81,08%	89,48%

Tabell 3

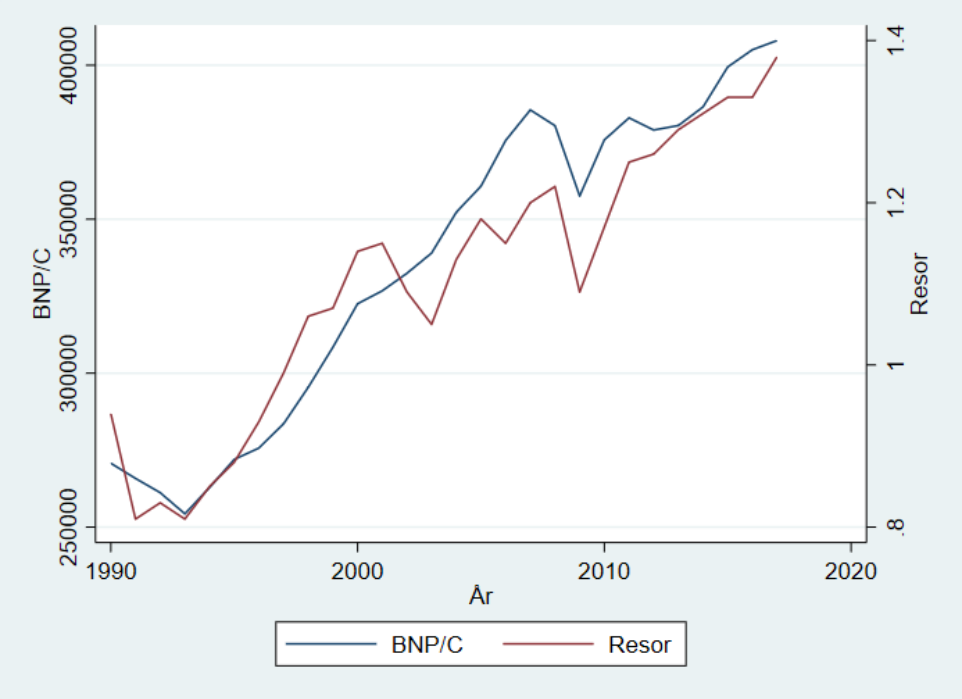
	BNP/capita		
	11300 (Poland)	34800 (Finland)	68200 (Norge)
Kurvans lutning vid 30 års ålder i procentenheter	-0,15799	-0,10074	-0,0521
Kurvans lutning vid 60 års ålder i procentenheter	-0,42625	-0,2901	-0,15857
Kurvans lutning vid 85 års ålder i procentenheter	-0,77802	-0,59951	-0,3678

Tabell 4

Inlämningsuppgift 2C

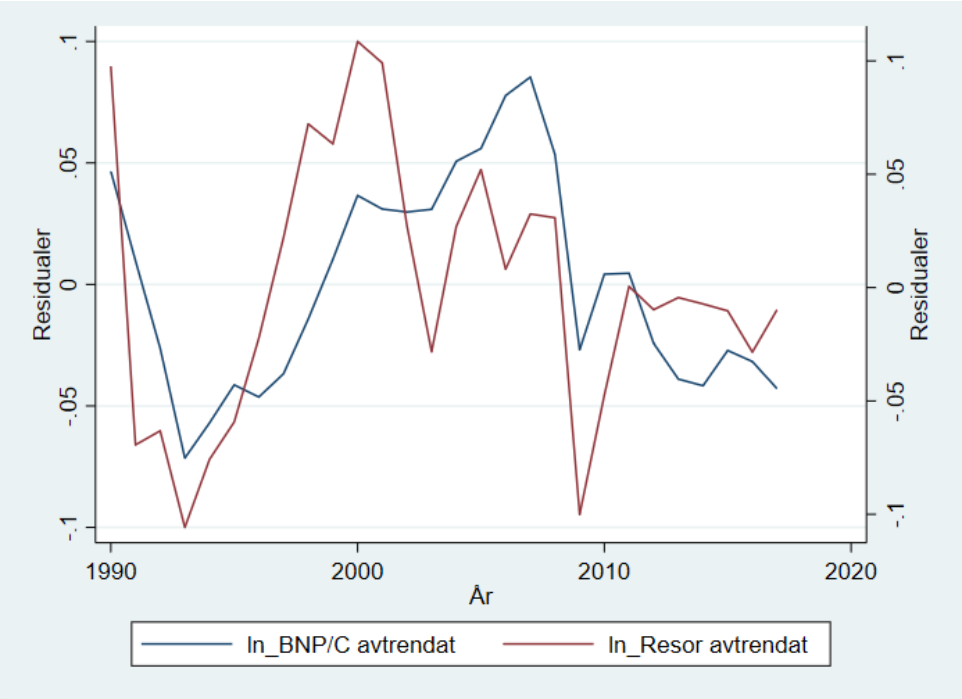
I följande rapport görs en analys av medelantalet flygresor för svenska medborgare och hur antalet ökar när Sveriges BNP/C ökar med tre procent. För att åstadkomma detta har data från år 1990 till 2017 med variablerna, antalet flygresor per svensk invånare och BNP/C använts.

Vidare, då komplexiteten av att analysera tidsseries ligger i att observationerna ofta är beroende av varandra och det krävs en förståelse för hur positiva trender påverkar både medelvärdet av antalet resor, och BNP/C positivt när tiden ökar. Detta kan ses genom graf 1 som mäter ökningen av våra variabler genom tiden. Ett annat sätt att se på det är att graf 1 förklarar en gemensam trend.



Graf 6, graf som visar hur BNP/C och Resor har förändrats övertid

Det som bör göras då, är att ta ur trenden ur våra två variabler genom att spara ner residualerna ifrån BNP/C och år, samt residualerna ifrån resor och år. Vi logaritmerar sedan resudialerna för att senare lättare kunna ta reda på hur mycket resandet ökar när BNP/C i procent. Resultatet är två nya variabler ln_BNP/C avtrendad och ln_Resor avtrendad som vi ser i graf 2. Vi kan se i grafen att de avtrendade variablerna tycks fortfarande samvariera efter bortrensningen av trenden.



Graf 7, graf som visar hur BNP/C och Resor har förändrats runt trenden

Problematiken här blir då istället att det mänskliga ögat inte är särskilt bra på att se autokorrelationen som dessa variabler har rent grafiskt. Vi kan inte helt lätt se om det finns en chans för randomwalk i vår data, även om vi rent teoretiskt vet att det antagligen inte borde vara det. Det är helt enkelt bättre att låta datorn räkna på korrelationskoefficienten i våra två variabler för att se om vi behöver differentiera dem. Detta görs genom pwcorr kommandot där vi mäter våra variabler emot sig själva fast med 'laggade' värden, det vill säga, vi mäter värde i variablerna med det föregående värdet i tidsserien. Resultatet blir att vi kan se att residualerna ifrån det logaritmerade värdet av BNP/C fortfarande har en hög autokorrelationskoefficient på närmare 0,90 och vår andra variabel för resor, även om den är betydligt mindre, får en autokorrelationkoefficient på närmare 0,56 vilket ändå är relativt högt.

Vi ställs då inför två dilemman i val av modell. Förvisso har vi starka empiriska bevis för att vi bör använda (d.) kommandot i Stata för att differentiera våra variabler, men vi har rätt starka teoretiska bevis för att BNP/C inte har ett perfekt minne. Vidare, med tanke på att vi bara har data mellan åren 1990 till 2017, så har vi betydligt mindre observationer än vad som rekommenderas för att använda Newey-West standardfel även om vi har relativt höga värden på våra autokorrelationskoefficienter.

Vi bestämmer oss slutligen för att använda modellen $\text{res\_ln(resor)} = \beta 0 + \beta 1 * \text{res\_ln(BNP/C)} + u$ utan Newey-West standardfel (Tabell 1, kol 2). Detta då det upplevs som att de teoretiska bevisen att BNP/C inte är en randomwalk är starkare än de empiriska bevisen vi har för att differentiera våra variabler. Vidare, som man ser i tabell 1, får en därmed underdrivna standardfel i vår modell, men med tanke på att vi har så få observationer hade Newey-West metoden ändå inte justerat standardfelen korrekt. Detta kan även ses i tabell 1, kol 3 och 4, där modellerna med Newey-West standardfel inte skiljer sig markant från de utan.

Slutligen, enligt vår modell (Tabell 1, kol 2), leder en ökning av 3 procent av GDP/C till en ökning på 2,4% i antalet resor per invånare i Sverige. Hade vi valt att differentiera våra variabler och istället använt modellen $\Delta\text{res\_ln(resor)} = \beta 0 + \Delta\beta 1 * \text{res\_ln(BNP/C)} + u$(tabell 1, kol 1), så

hade en 3 procentig ökning av GDP/C lett till en 3,975% ökning av antalet resor per invånare i Sverige.

VARIABLER	(1) Regression med differentiering	(2) Regression utan differentiering	(3) Regression med differentiering och Newey-West standardfel	(4) Regression utan differentiering och Newey-West standardfel
D.res_ln_BNP/C	1.325**		1.325**	
	(0.329)		(0.280)	
res_ln_BNP/C		0.800**		0.800**
		(0.187)		(0.220)
Konstant	0.000415	-1.42e-10	0.000415	-1.42e-10
	(0.00785)	(0.00896)	(0.00775)	(0.0111)
Observationer	27	28	27	28
R-squared	0.407	0.358		
Robusta standardfel inom paranteserna				
** p<0.01, * p<0.05				

Tabell 5