Gretl

Gnu Regression, Econometrics and Time-series

En introduktion

Lars Pålsson Syll
Malmö högskola
April 2009

Innehåll

6. Några grundläggande sannolikhetsbegrepp

7. Diagram, tabeller, centralvärde och spridningsmått

7.1 Stapeldiagram
7.2 Histogram
7.3 Sambandsdiagram
7.4 Tidsseriediagram
7.5 Andra typer av diagram
7.6 Ett skolexempel
7.7 Kort om icke-linearitet
7.8 Frekvenstabeller
7.9 Centralvärden
7.10 Spridningsmått

1 Introduktion

I det här kapitlet kommer du att få stifta bekantskap med några av de mer grundläggande egenskaperna hos **gretl**. Du kommer bl a att lära dig hur man installerar programmet, hur man använder de olika fönstren och hur man importerar data.

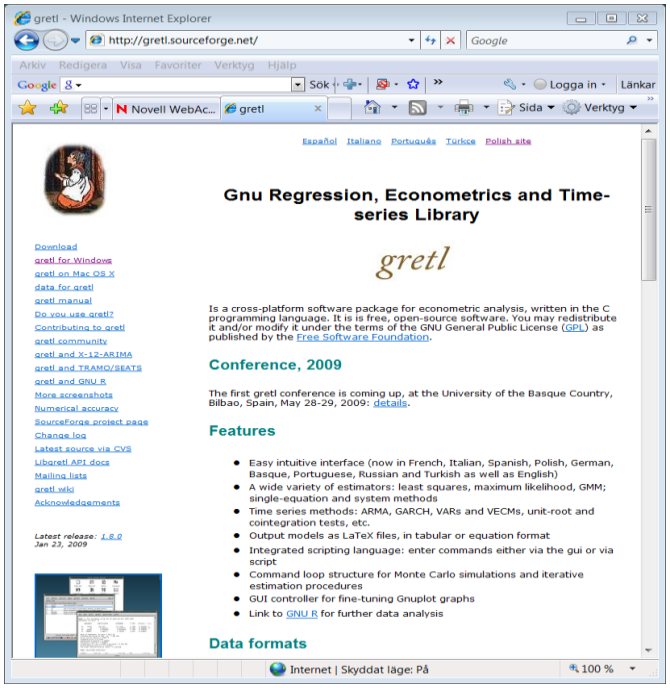
1.1 Vad är gretl?

Gretl är en akronym för Gnu Regression, Econometrics and Time-Series Library. Det är ett statistikprogram som både är mycket användarvänligt och kraftfullt för att utföra en rad statistiska och ekonometriska beräkningar. Och det bästa av allt – det är gratis! Det kan laddas ner från http://**gretl**.sourceforge.net och installeras på din dator.

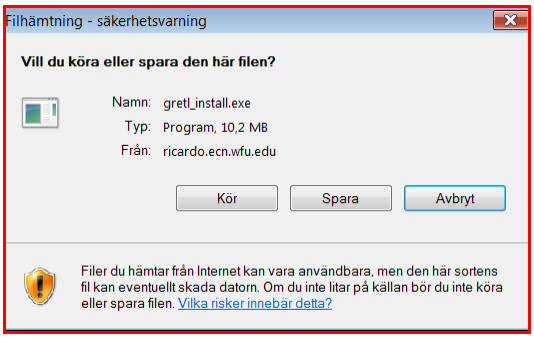
Gretl kommer med en stor uppsättning datafiler och en rejäl databas (med företrädesvis amerikanska makroekonomiska tidsserier). Från **gretl**s webbsida kan man också ladda ner ett stort urval datafiler från ledande läroböcker i ekonometri. **Gretl** kan användas för att exempelvis beräkna minsta kvadratskattningar (OLS), icke-linjära (NLS) och viktade kvadratskattningar (WLS). **Gretl** använder ett separat program – gnuplot – för att skapa grafer och diagram.

1.1.1 Hur man installerar gretl

För att installera **gretl** om du använder Microsoft Windows går du till http://**gretl**.sourceforge.net/win32/ (använder du Macintosh, Linux eller någon annan vanlig plattform finns det versioner för dessa också):

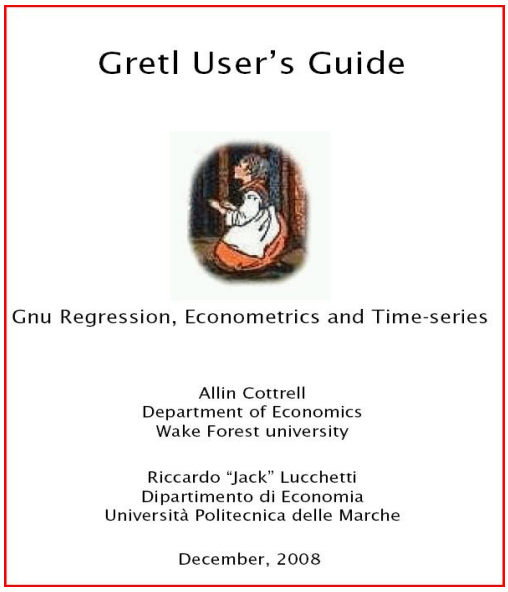


Om du inte själv vill kompilera en egen version av **gretl** är det lättast att bara installera den självinstallerande filen **gretl** install.exe som du hittar genom att trycka på **gretl** for windows ute till vänster på hemsidan och sedan finner under Download. Tryck på **gretl** install.exe och du får fram



Här trycker du på knappen Kör och installerar programmet på din dator.

Med **gretl** kommer också en manual i Adobe pdf format där man mer i detalj bland annat kan läsa om hur man installerar och använder gränssnittet.



1.1.2 Grundläggande egenskaper i gretl

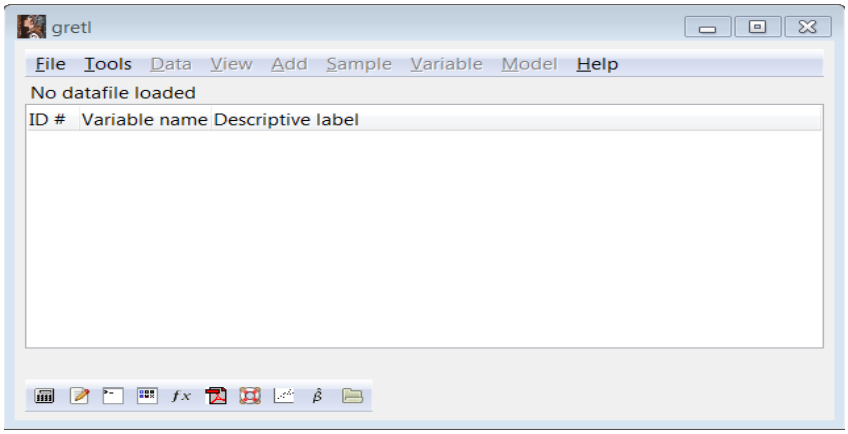
Det finns flera olika sätt att arbeta med **gretl**. Det har ett mycket enkelt och intuitivt språk, men vi kommer främst att här fokusera på användning av det grafiska användargränssnittet (GUI). Detta fungerar ungefär som i t ex MS Windows och Excel. Du använder datormusen för att klicka på knappar som öppnar upp olika dialogrutor. Där klickar du i de önskade valen och utför kommandona genom att trycka på OK-knappen.

I **gretl** finns också möjlighet att använda ett kommandolinjegränssnitt. Detta kan ske antingen från konsolen (**console**) eller genom att använda skript (**scripts**). Utöver dessa finns också möjlighet att helt skippa dialog och använda gretlcli direkt i ett dos-kommandofönster (vilket kanske främst är av intresse för Linux-användare). Vi kommer inte att använda detta i den här manualen.

Det som de flesta antagligen kommer att tycka är det smidigaste sättet att exekvera enskilda **gretl**kommandon med är **gretl**s **console.** Hur man använder den kommer vi att gå igenom i sektion 1.3.1.

Vill man däremot exekvera en hel serie kommandon är det bäst att göra detta via **scripts**. En av de många finesserna i **gretl** är att alla de enskilda kommandon som exekveras från konsolen sparas i en kommandolog (**command log**). Därför kan man lätt köra alla kommandona i ett vid ett senare tillfälle genom att kopiera kommandologen i **scripts** och exekvera dem därifrån. Detta behandlas i sektion 1.3.2.

I figuren nedan finner du huvudfönstret i **gretl.**



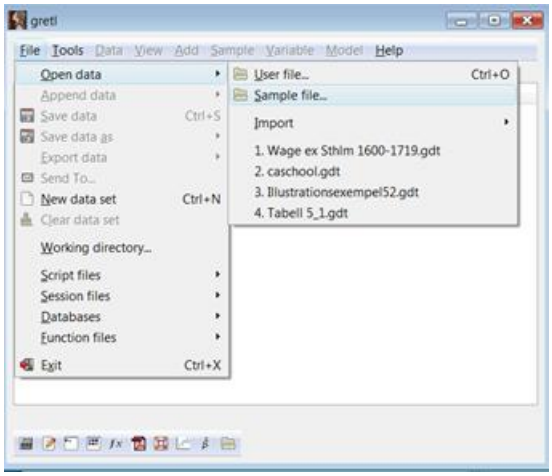
Längst upp i fönstret finner du menyfältet. Härifrån kan du importera och manipulera data, analysera data och styra hur resultaten ska presenteras.

Längst ner i fönstret finner du **gretl**s verktygsfält med en rad nyttiga funktioner som vi ska återkomma till.

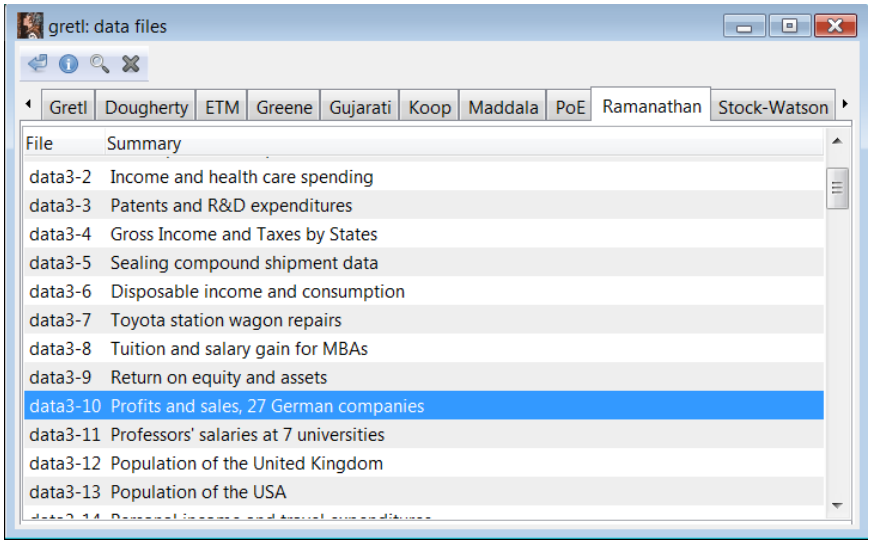
1.2 Hur man importerar data

Gretl erbjuder stora möjligheter att importera en rad olika filformat. Och när man väl öppnat upp dem i **gretl** kan de sedan vid behov exporteras i en rad olika filformat.

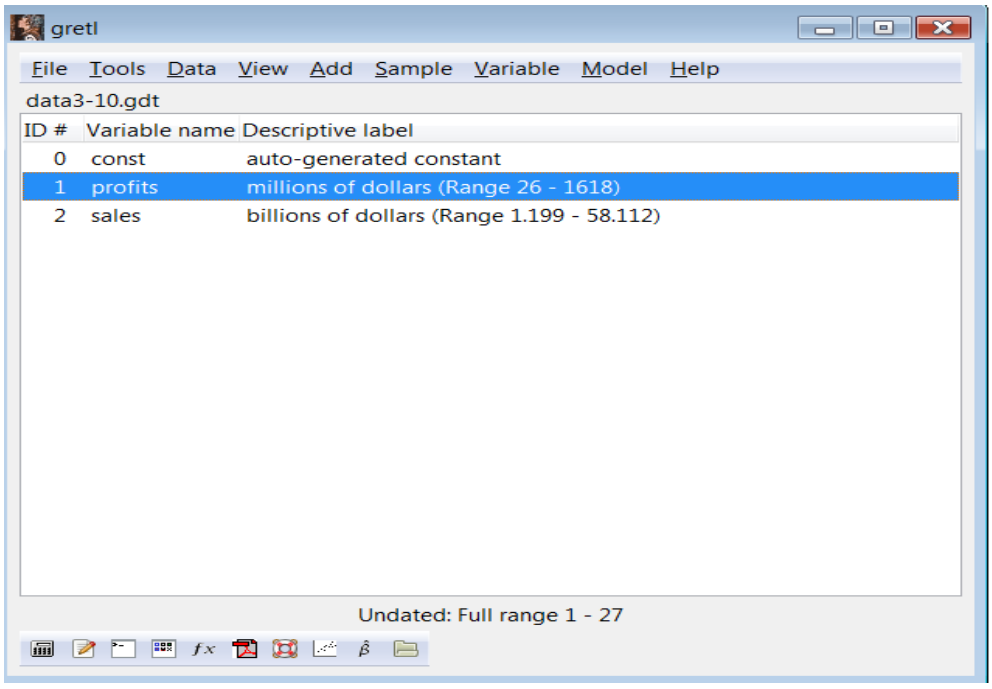
Vi laddar ner genom att klicka på File Open data > Sample file som i figuren nedan.



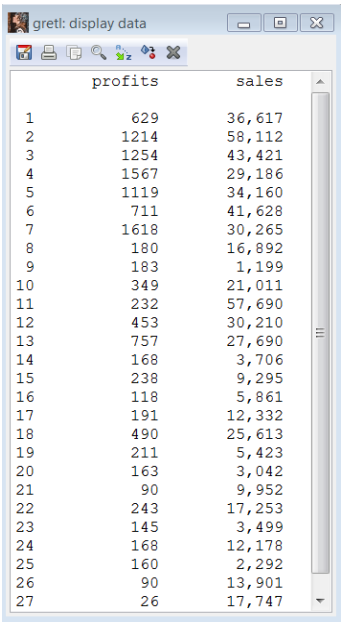
Ett annat sätt att ladda ner data på, är att trycka på knappen för att öppna en datafil i verktygsfältet. Knappen ser ut som en folder och finns längst till höger i verktygsfältet. Då öppnas ett annat fönster upp med knappar för de olika dataset som du installerat i **gretl**/data directory i ditt program.



Klicka på Ramanathan och rulla ner tills du hittar datafilen data3-10.
Markera med datormusen och dubbelklicka. Detta laddar ner filen i **gretl**.



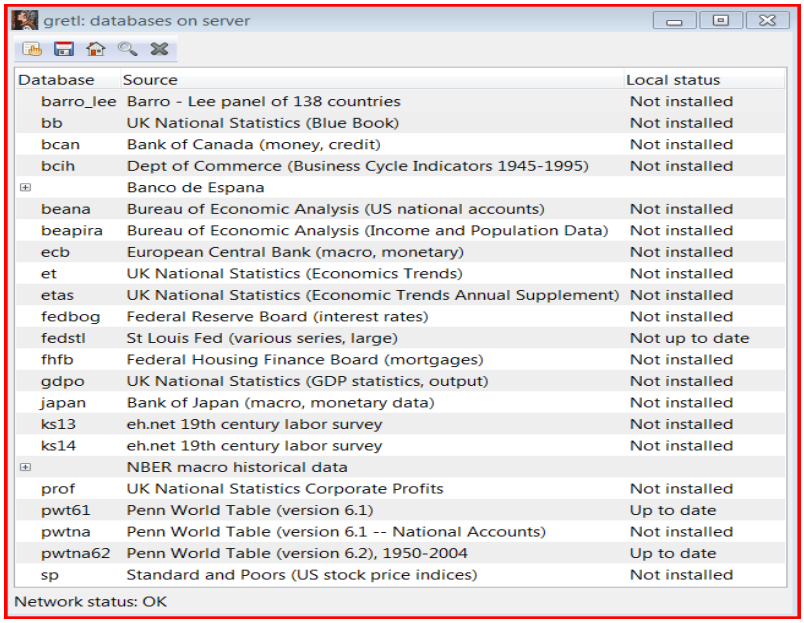
Markera de båda variablerna profits och sales. Klickar du sedan på Data i menyfältet och i rullgardinsmenyn på Display values får man värdena som i figuren till höger.


gretl: display data

	profits	sales
1	629	36,617
2	1214	58,112
3	1254	43,421
4	1567	29,186
5	1119	34,160
6	711	41,628
7	1618	30,265
8	180	16,892
9	183	1,199
10	349	21,011
11	232	57,690
12	453	30,210
13	757	27,690
14	168	3,706
15	238	9,295
16	118	5,861
17	191	12,332
18	490	25,613
19	211	5,423
20	163	3,042
21	90	9,952
22	243	17,253
23	145	3,499
24	168	12,178
25	160	2,292
26	90	13,901
27	26	17,747

Från rullgardinsmenyn kan du bland annat editera eller lägga till observationer och bestämma datafilens struktur. De alternativ som finns är tidsserie, tvärsnitt och panel. De alternativ **gretl** ger dig beror på vilken struktur du har angivit att datan har. Om ett **gretl**kommando inte är tillgängligt för den definierade datastrukturen blir den gråfärgad i rullgardinsmenyn.

Notera i figuren att **gretl** ger dig möjlighet att importera data. Via File > Open data > Import kan du se vilka dataformat som **gretl** kan importera (ASCII, CSV, EXCEL m fl). Från File rullgardinen kan du också se vilka olika format du kan exportera **gretl**filer i. Via File > Databases > On database server kan du nå en rad dataset som du kan ladda ner i **gretl** på samma sätt som vi beskrev ovan.


gretl: databases on server

Database	Source	Local status
barro_lee	Barro - Lee panel of 138 countries	Not installed
bb	UK National Statistics (Blue Book)	Not installed
bcan	Bank of Canada (money, credit)	Not installed
bcih	Dept of Commerce (Business Cycle Indicators 1945-1995)	Not installed
⊞	Banco de Espana	
beana	Bureau of Economic Analysis (US national accounts)	Not installed
beapira	Bureau of Economic Analysis (Income and Population Data)	Not installed
ecb	European Central Bank (macro, monetary)	Not installed
et	UK National Statistics (Economics Trends)	Not installed
etas	UK National Statistics (Economic Trends Annual Supplement)	Not installed
fedbog	Federal Reserve Board (interest rates)	Not installed
fedstl	St Louis Fed (various series, large)	Not up to date
fhfb	Federal Housing Finance Board (mortgages)	Not installed
gdpo	UK National Statistics (GDP statistics, output)	Not installed
japan	Bank of Japan (macro, monetary data)	Not installed
ks13	eh.net 19th century labor survey	Not installed
ks14	eh.net 19th century labor survey	Not installed
⊞	NBER macro historical data	
prof	UK National Statistics Corporate Profits	Not installed
pwt61	Penn World Table (version 6.1)	Up to date
pwtna	Penn World Table (version 6.1 -- National Accounts)	Not installed
pwtna62	Penn World Table (version 6.2), 1950-2004	Up to date
sp	Standard and Poors (US stock price indices)	Not installed

Network status: OK

1.3 Att använda gretls språk

Även om **gretl**s GUI är suveränt lätt att använda kan det ibland vara fördelaktigt och snabbare att använda **gretl**s språk. Detta sker antingen via console eller script.

Kom ihåg att språket är beroende av om man använder versaler eller kapital. Det innebär att y och Y är två helt olika saker. Det gäller därför att vara observant på detta när man exempelvis skriver kommandon eller anger en variabel.

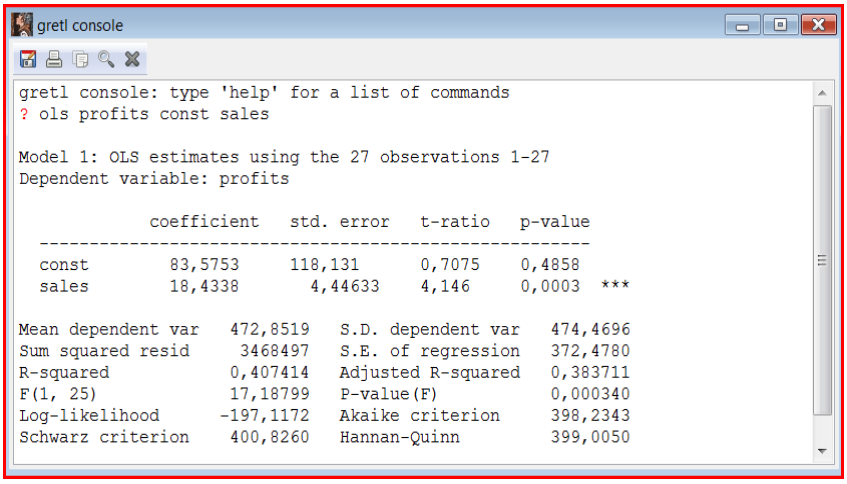
1.3.1 Console

Med konsolen kan man exekvera kommandon interaktivt i **gretl**. Om man trycker på console-knappen i verktygsfältet (eller från rullgardinen som man får fram om man trycker på Tools) får man fram ett nytt fönster där man efter markören ”?” radvis kan exekvera sina kommandon.

Låt oss ta ett exempel. Vid markören ”?” kan du skatta en modell genom att använda minsta kvadratmetoden (OLS) genom att skriva

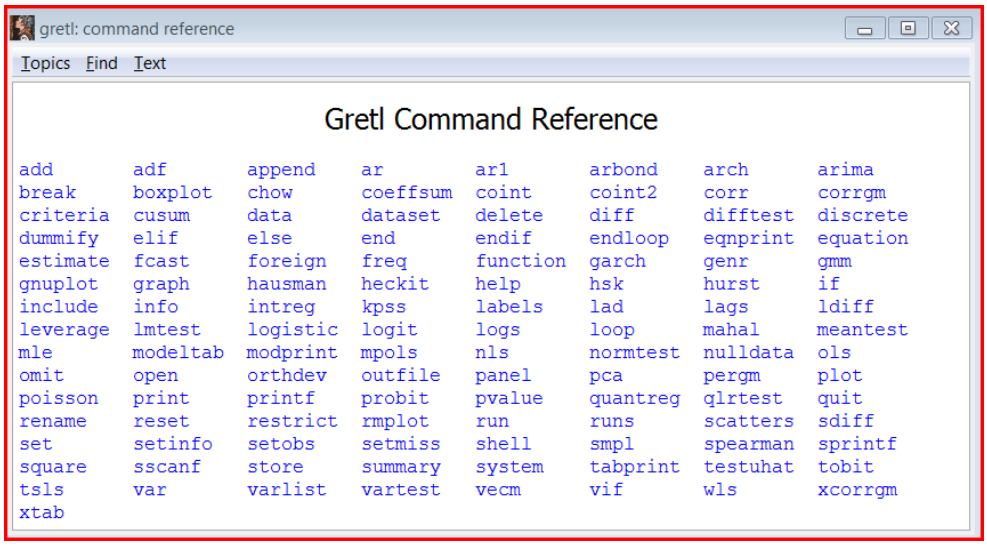
ols profits const sales

Resultatet dyker upp som output i konsolfönstret.



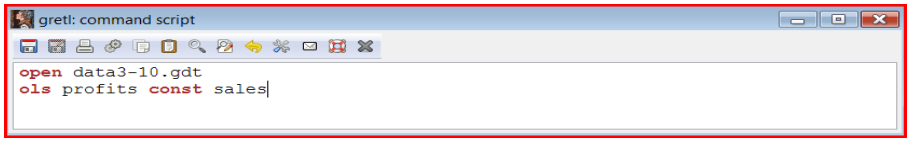
Kommandona sparas och kan återkallas genom att man helt enkelt med upp-tangenten rullar genom de tidigare kommandona tills man kommer till det man vill använda. Väl här kan man ändra och anpassa kommandot och rätta till eventuella syntaxfel innan man slår på returtangenten för att exekvera kommandot.

För att utföra kommandona är det nödvändigt att man känner till språksyntaxen. Den finner man i **gretl**s command reference, som du kan nå genom att i verktygsfältet trycka på knappen som ser ut som en livboj (fjärde från höger). Du kan också nå den genom att trycka på Help eller tangentbordets genväg F1.



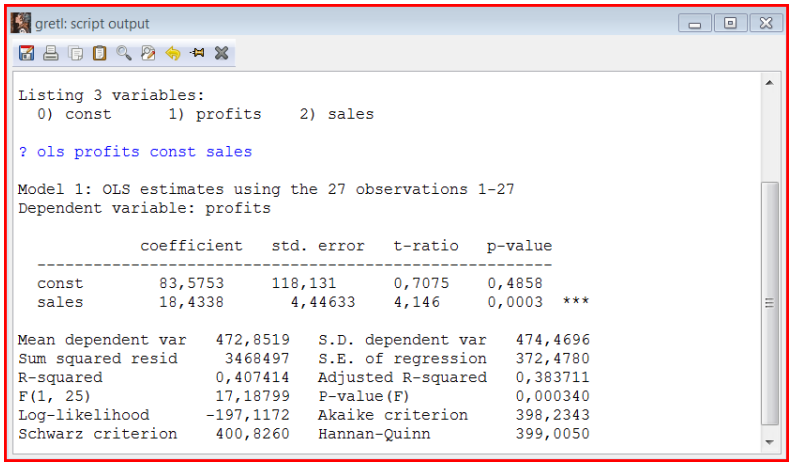
1.3.2 Scripts

Gretlkommandon kan också som vi nämnt samlas i en fil och exekveras på en gång. Man öppnar ett nytt command script från filmenyn. I rullgardinsmenyn öppnar File > Script files > New script upp en editor för kommandoskript. Kommandona skrivs in på så vis att man använder en rad för varje kommando. Räcker inte en rad till kan man använda ”\” för att fortsätta kommandot på flera rader. I vårt exempel ser det ut som i figuren nedan.



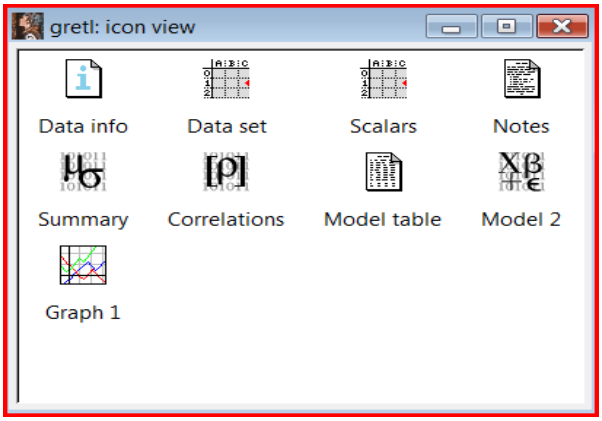
För att spara filen trycker man på sparfliken längst upp till vänster.

För att köra programmet klickar man på den kugghjulsliknande knappen. Outputen dyker då upp i ett separat fönster.



1.3.3 Sessioner

Gretl använder sig av ett sessionskoncept för att göra det möjligt för dig att spara modeller, grafer, datafiler, kommentarer och dylikt i ett gemensamt utrymme. I sessionsfönstret ligger de olika objekten samlade som ikoner som kan sparas för att användas senare. När du sparar din session ska de objekt du lagt till vara tillgängliga nästa gång du öppnar upp sessionen.



För att lägga till en modell till din session använder du File > Save to session as icon från modellens rullgardinsmeny. Vill man spara en graf, högerklickar man på grafen och väljer save to session as icon.

Glöm inte att spara sessionen om du vill kunna gå tillbaka och se på dina körningar! I **gretl**s huvudfönster väljer du File > Session files > Save session.

När du sparat en modell eller graf kommer dess ikon att dyka upp i fönstret session icon view. Om du dubbelklickar på ikonen kommer objektet att visas i ett nytt fönster.

2 Enkel linjär regression

2.1 En enkel linjär regressionsmodell

Den vanliga enkla regressionsmodellen är

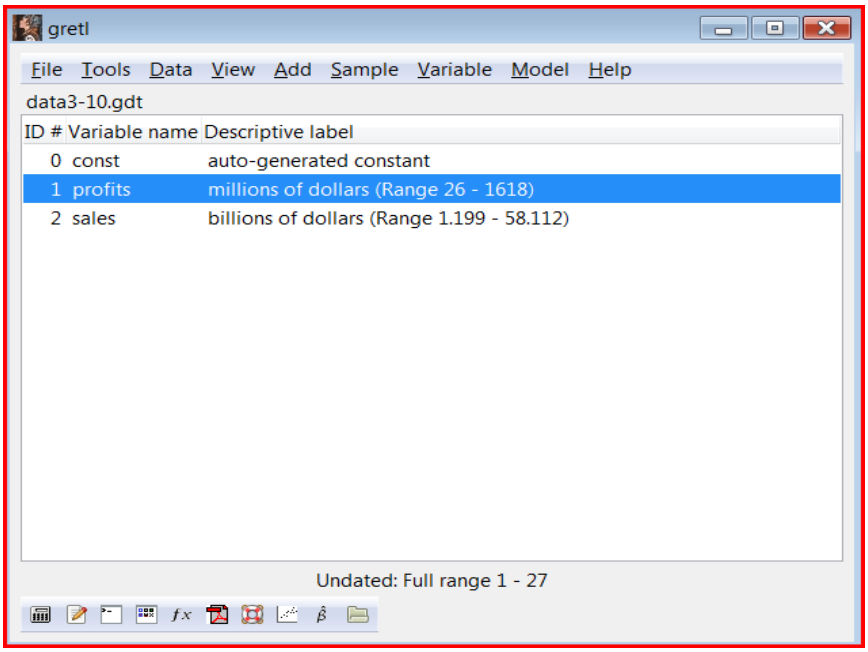
$$y_t = \alpha_1 + \beta_2 x_t + e_t \quad t = 1, 2, ..., T$$

där y_t är den beroende variabeln, x_t den oberoende variabeln och e_t är residualer och β_1 och β_2 är de parametrar som du vill skatta.

2.2 Hur man öppnar upp datan

Det första du ska göra är att ladda ner datan i **gretl**.

Ladda ner data 3-10.gdt genom kommandona File > Open data > sample file från menyraden och gå in under mappen Ramanathan. Välj sedan att markera variabeln profits.

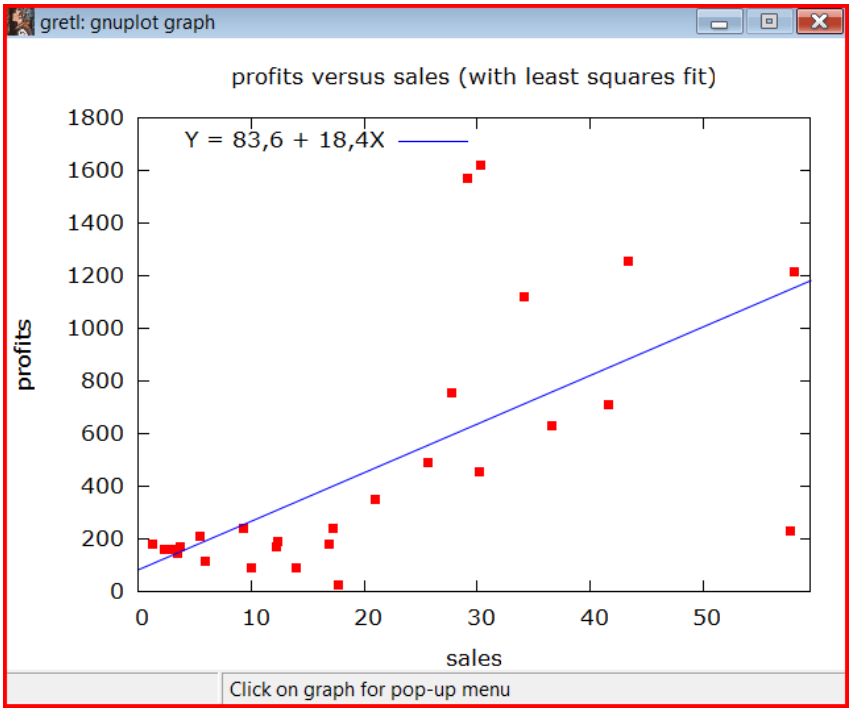


I kolumnen under Descriptive label finner du (eventuellt) information om de variabler som finns i programminnet. Om de saknas kan man själv lägga till etiketter genom att markera variabeln och sedan högerklicka på den. Då får man fram en rullgardinsmeny. I den dubbelklickar man på Edit attributes. Då öppnas en dialogruta upp där man exempelvis kan beskriva variabeln eller ändra dess namn. Man kan alternativt också via huvudfönstret välja Variable > Edit attributes.

2.3 Skapa en graf

För att skapa en graf kan man i verktygsfältet klicka på knappen som ser ut som en graf (tredje från höger). I dialogrutan som poppar upp kan du då ange vilka variabler du vill plotta mot varandra

Om vi använder den tidigare filen data 3-10.gdt får vi om vi markerar de två variablerna sales och profits följande graf om vi anger variabeln sales på x-axeln och profits på y-axeln (vilket verkar rimligt med tanke på att vi väl antar att sales är den oberoende variabeln och profits den beroende). Hur man kan gå in i grafen och redigera återkommer vi till senare.

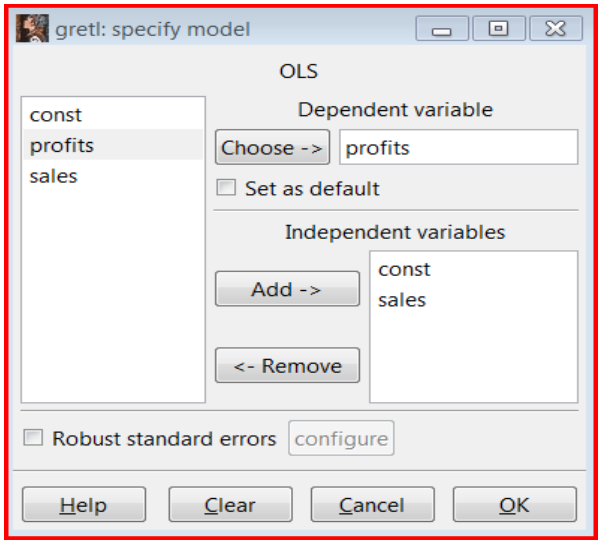


2.4 Hur man skattar relationen mellan sales och profits

Låt oss nu skatta parametrarna för profitekvationen

$$profits_t = \alpha_1 + \beta_2 sales_t + e_t \quad t = 1, 2, ..., T$$

Från menybaren väljer du Model>Ordinary Least Squares i rullgardinsmenyn för att öppna upp följande dialogruta.

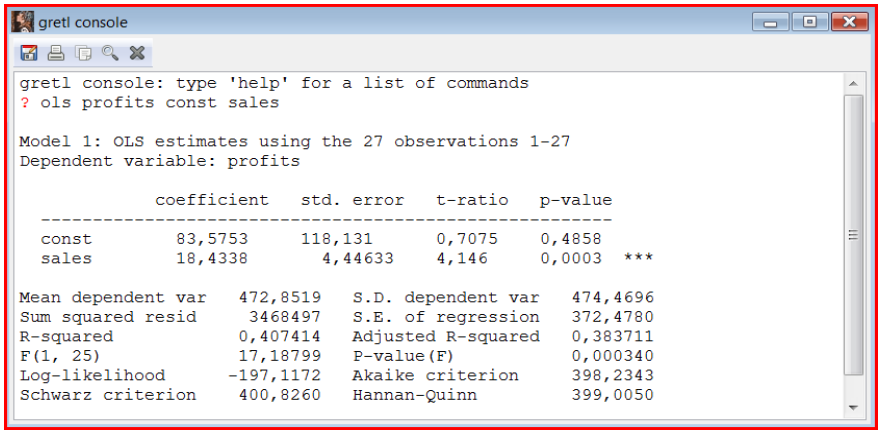


Här talar du om för **gretl** vilken variabel som ska vara beroende och vilken som ska vara oberoende. Du behöver aldrig ange att en konstant ska ingå i modellen, eftersom **gretl** i standardinställningen alltid inkluderar en konstant genom att automatiskt placera variabeln const i listan. För att inkludera profits som oberoende variabel markerar man den och klickar på ”Add→” knappen.

Alternativt kan man via konsolen efter ”?” skriva in

```
ols profits const sales
```

för att skatta regressionsfunktionen. Syntaxen är enkel. ols talar om för **gretl** att du vill göra en minsta kvadratskattning (där den första variabeln är den beroende och variabeln efter const är den oberoende). Resultatet blir (som vi tidigare sett):



2.5 Prediktioner

Gretl kan självklart användas för att göra prediktioner. Vår regression gav resultatet

$$profits_t = 83.5753 + 18.4338sales_t$$

Om vi då vill ha reda på hur stor vinsten blir för ett genomsnittligt företag vid ett försäljningsvärde på 20 (miljarder $) får vi

$$profits_t = 83.5753 + 18.4338(20) = 452.25.$$

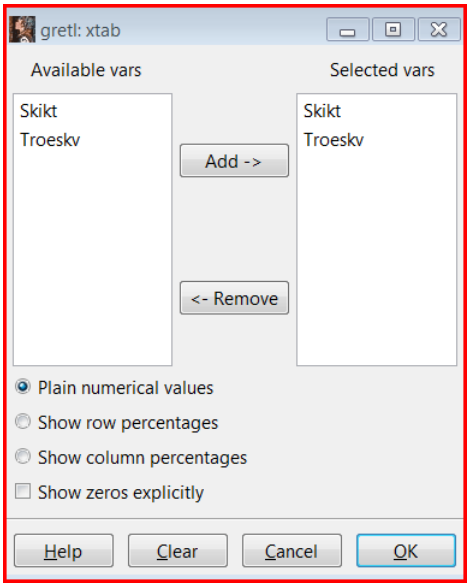
Skriptet i **gretl** för att få detta är

```
genr yhat = $coeff(const) + $coeff(sales)*20
```

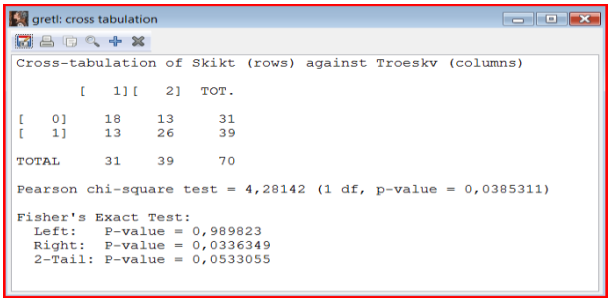
2.6 Olika typer av sambandsmått

En vanlig typ av sambandsmått för nominalskalemätta variabler är χ^2-testet. Det diskuteras i kapitel 7 i *Eggeby & Söderberg* och används till exempel när vi har skapat korstabeller utifrån grupperade observationer i kvalitativa kategorier.

Låt oss som exempel ta datafilen Tabell 7_1Xsqtest.gdt (som du finner tillsammans med andra datafiler i mappen N:\LUT\STUD\HISTLIC). Här har vi två kategorivariabler Skikt (0=torpare, 1=bönder) och Troeskverk (2=utan tröskverk, 1=med tröskverk). I huvudfönstret väljer vi View>Cross Tabulation. För över variablerna Skikt och Troeskverk till rutan ”Selected vars”:

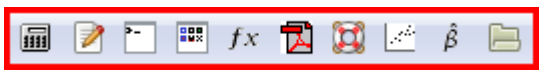


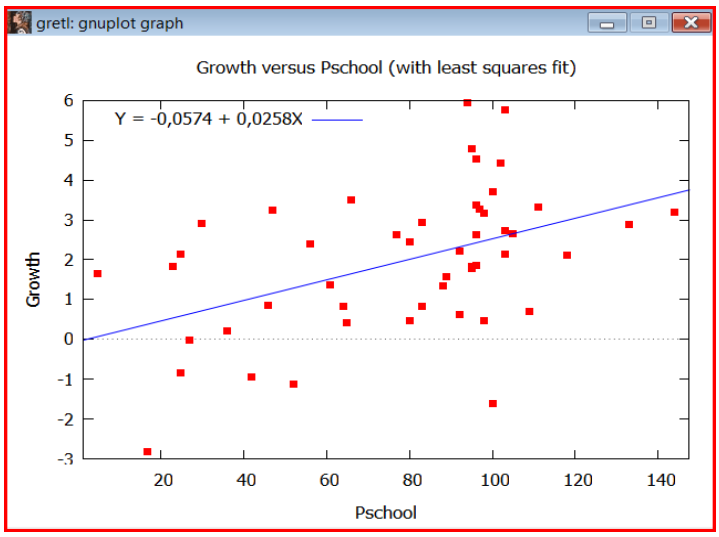
Klicka på OK och du får följande:



Värdet på χ^2 är 4.28 (vilket är större än det kritiska χ^2-värdet på 3.84, som du kan få fram via Tools>Statistical tables>chi-square) och det tyder på att det föreligger ett signifikant samband mellan social ställning och tröskverksinnehav. Vi ser också att p-värdet – som anger sannolikheten för att resultatet skulle vara slumpmässigt – bara är 0.038. Sambandet är signifikant.

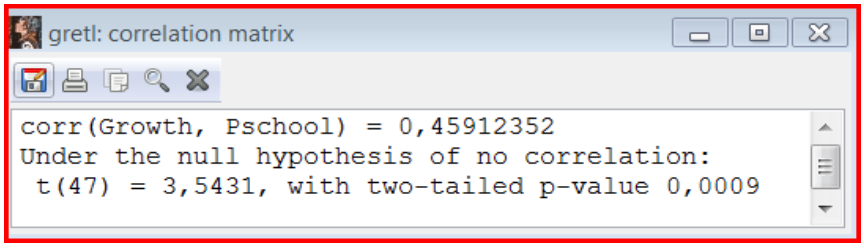
För att se om det föreligger ett statistiskt samband mellan olika variabler är det bra att använda sig av olika slags diagram. Låt oss exempelvis se på relationen mellan tillväxt och ett antal andra faktorer. Ladda ner datafilen Tabell 5_1.gdt. Markera Pschool (som anger skolgångens omfattning) och Growth (som anger tillväxten i ekonomin) och högerklicka sedan på någon av dem och välj XY Scatterplot, eller klicka på tredje symbolen från höger längst ner i fönstret:



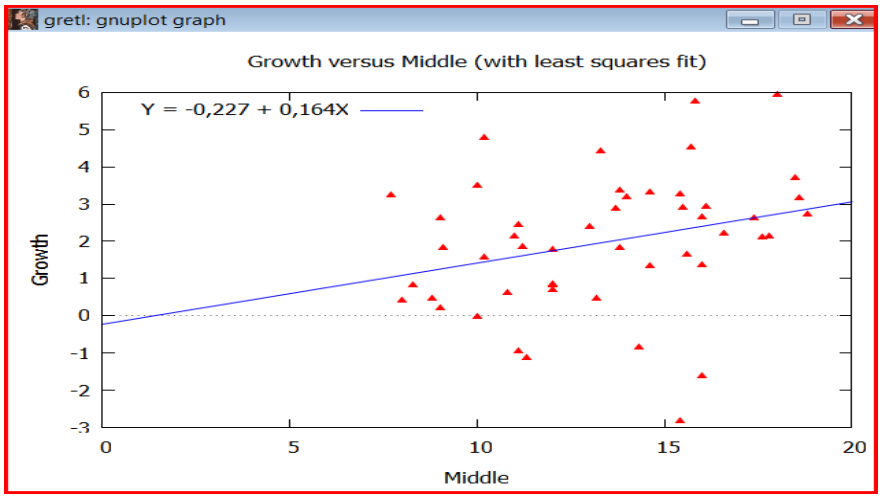


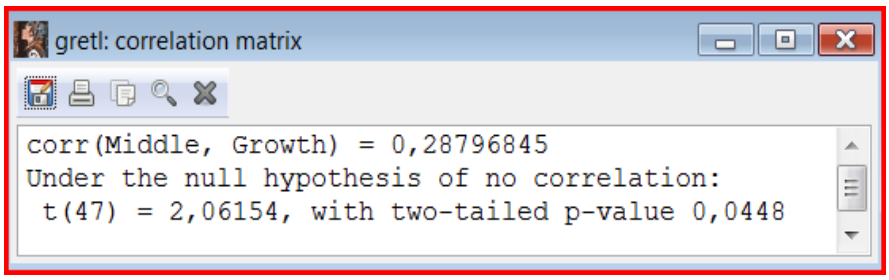
och välj Pschool som x-axelvariabel. Du får då följande diagram:

I prickdiagrammet läggs automatiskt in en minstakvadrat-skattning av sambandet. Väljer vi Correlation matrix i stället för XY scatterplot får vi:



Som vi ser är korrelationen mellan variablerna rätt stark, 0.46. Gör vi motsvarande för Middle (som anger jämlikhet i inkomst-fördelningen) och Growth får vi följande diagram och fönster:



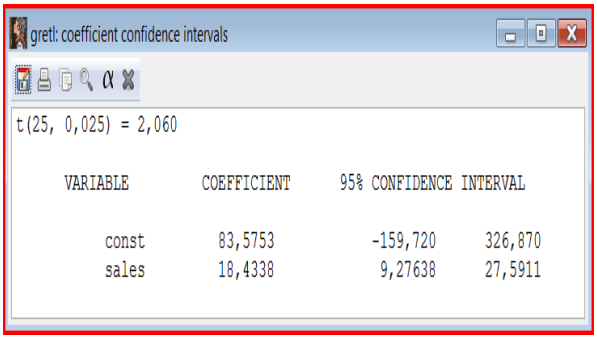

gretl: correlation matrix
corr(Middle, Growth) = 0,28796845
Under the null hypothesis of no correlation:
t(47) = 2,06154, with two-tailed p-value 0,0448

3 Konfidensintervall och hypotesprövning

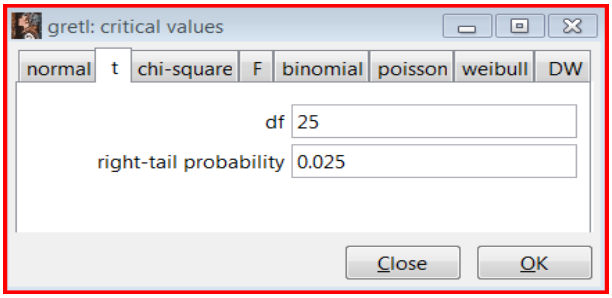
I det här kapitlet ska jag korta visa hur man i **gretl** kan skatta konfidensintervall och testa hypoteser. Eftersom **gretl** har flera mycket behändiga verktyg för detta är det lätt att erhålla kritiska värden och p-värden för de vanligaste sannolikhetsfördelningarna. De enklaste sätten att göra det är att antingen använda dialogrutorna eller **gretl**s programspråk.

3.1 Konfidensintervall

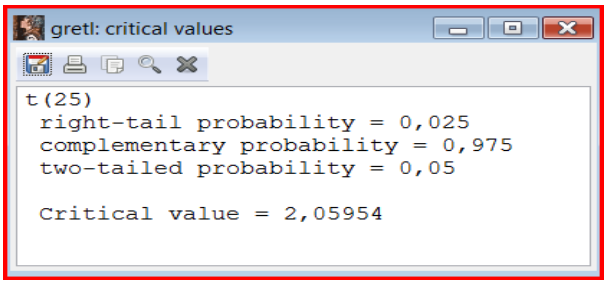
Det är alltid viktigt att ha ett mått på hur precisa ens parameterskattningar är.
Anta att vi har kört regressionen i förra kapitlet (ols profits 0 sales). Vi ser då i outputfönstret att standardavvikelsen för β_1 är 4.44633. Om vi klickar på Analysis>Confidence intervals for coefficients i modellfönstrets rullgardinsmeny kan vi se att det kritiska t-värdet t(25, 0.025) = 2.060 och att det 95% intervallet för β_2 är 9.27638 - 27.5911. Detta innebär att vi med 95 % säkerhet kan anta att β_2 har ett värde som ligger i detta intervall.



De kritiska t-värdena (notera att de anges som α/2) kan också fås via Tools>Statistical tables. Klicka på ”t” och skriv in värdena.



Klicka sedan på ”OK” och vi får följande output.



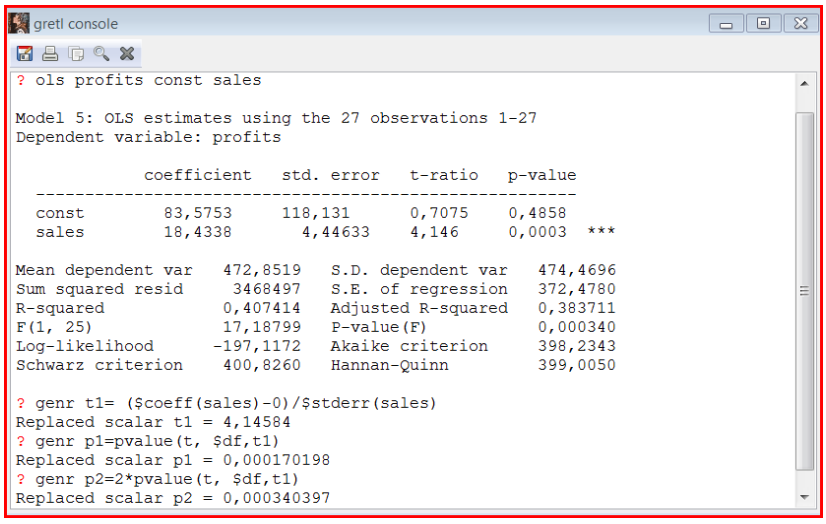
Självklart kan man också använda något av **gretl**s programkommandon om man vill. Syntaxen är

genr p = pvalue(distribution, parameters, xval).

För p-värdefunktionen skriver man då:

genr t1 = ($coeff(sales)-0)/$stderr(sales)
genr p1 = pvalue(t,$df,t1) (för enkelsidigt test)
genr p2 = 2*pvalue(t,$df,t1) (för att få tvåsidigt test).

I konsolen (självklart kan man också spara kommandona som ett skript som kan exekveras vid behov) ser resultatet ut så här:



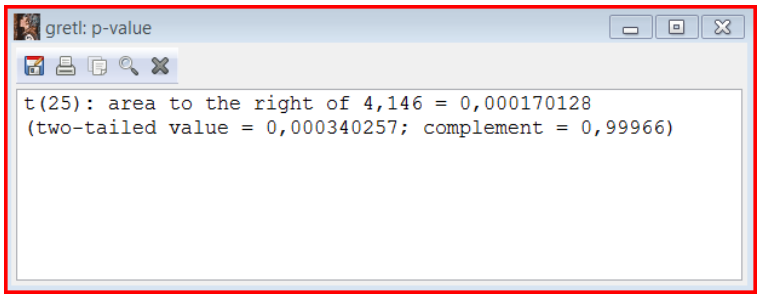
3.2 Hypotesprövning

Med hjälp av hypotestestning kan man jämföra ens tidigare uppfattningar om ett samband med vad datan visar. Exempelvis kan vi i vårt tidigare exempel testa hypotesen att $\beta_2 = 0$ mot alternativet att det är > 0. Teststatistiken är

$T = (b_2 - 0)/se(b_2) \sim t25$

givet att $\beta_2 = 0$ (nollhypotesen är sann). Välj $\alpha = 0.05$ vilket ger det kritiska värdet för den ensidiga mothypotesen ($\beta_2 > 0$), som är 1.708. Detta värde får du via rullgardinmenyn Tools > Statistical tables > t och där välja df = 25 och right-rail probability = 0.05. Beslutsregeln är att förkasta H_0 och föredra alternativet H_a om din t-statistik är större än 1.708. I modellfönstret kan man se att värdet för t (under t-ratio för variabeln sales) är 4.146. H_0 kan alltså förkastas (se också på p-värdet som är 0.00034).

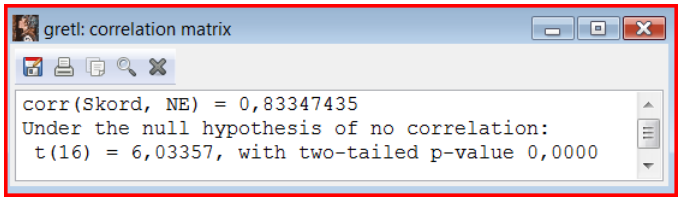
P-värdet kan man också få via Tools>P-value finder:



3.3 Regressionsanalys av skördeutfall

Eggeby & Söderberg tar i kapitel 8 upp enkel linjär regressionanalys. Med utgångspunkt från ett datamaterial från 18 byar i Luleå socken omkring 1850 försöker man ge svar på frågan om vilken effekt skörden per capita hade på nötkreatursinnehavet.

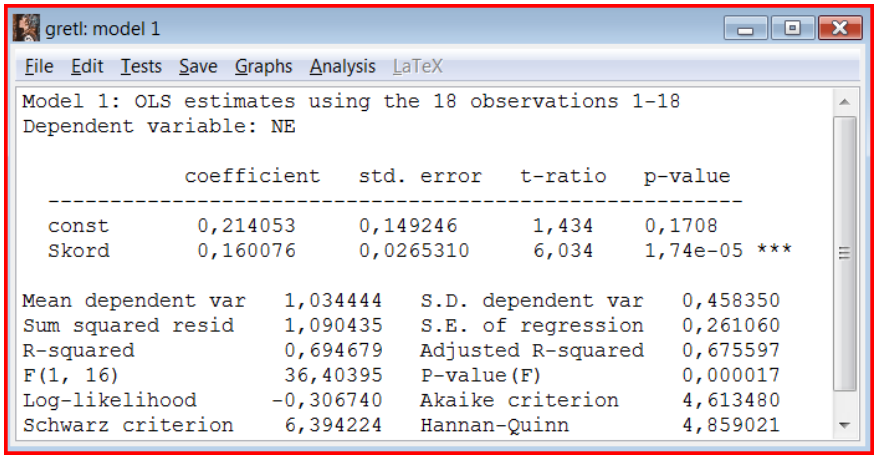
Rimligt verkar vara att anta att en större skörd skulle göra det möjligt att hålla ett större kreatursbestånd. Låt oss börja med att se på om det finns en korrelation mellan dessa variabler. Låt oss öppna datafilen Nedre Luledalens byar, markera variablerna NE och Skord och i rullgardinsmenyn välja correlation matrix. Vi får då följande figur:



Korrelationskoefficienten r_{xy} är 0.833. För att kolla om detta samband också kan vara kausalt testar vi modellen

ols NE const Skord

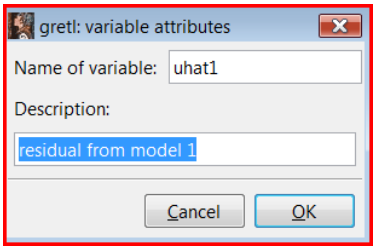
och får följande resultat:



Vi tolkar resultatet som att antalet nötkreatur/capita i genomsnitt ökar med 0.16 (nötkreaturs)enheter när skörden/capita ökar med en enhet (hektoliter).

Vi ser också att modellskattningen är ”statistiskt signifikant” (sannolikheten att parametervärdet för Skord = 0 understiger en promille). Förklaringsgraden r^2 är 0.69 (= 0.833^2), vilket innebär att cirka 70% av variationen i nötkreatursenheter/cap kan hänföras till variationen i skörden.

För att testa om detta är en ”bra” modell kan man titta på residualerna - skillnaderna mellan de verkliga observerade värdena och modellens skattningar av dessa värden. I modellfönstret klickar man Save>Residuals och får då fram följande fönster. Klicka på OK. Residualerna läggs till som ytterligare en variabel, uhat1.



4 Multipel regression

Den multipla regressionsmodellen är till sin struktur i grunden som den enkla regressionsmodellen. Skillnaden är att vi i stället för *en* oberoende variabel har *flera* oberoende variabler. Tolkningen av koefficienterna blir också något annorlunda. Den generella formen av modellen är

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_K x_{iK} + e_i \quad = 1, 2, \ldots, N \qquad (5.1)$$

där y_i är den beroende variabeln, x_{ik} den i:te observationen av den k:te oberoende variabeln, k = 2, 3, …,K, e_i är residualen och β_1, β_2,…,β_K är de parametrar vi vill skatta. Precis som i den enkla regressionsmodellen har residualen ett genomsnittsvärde på noll för varje värde på de oberoende variablerna. Varje residual har också samma varians,σ^2, och de är icke-korrelerade med varandra.

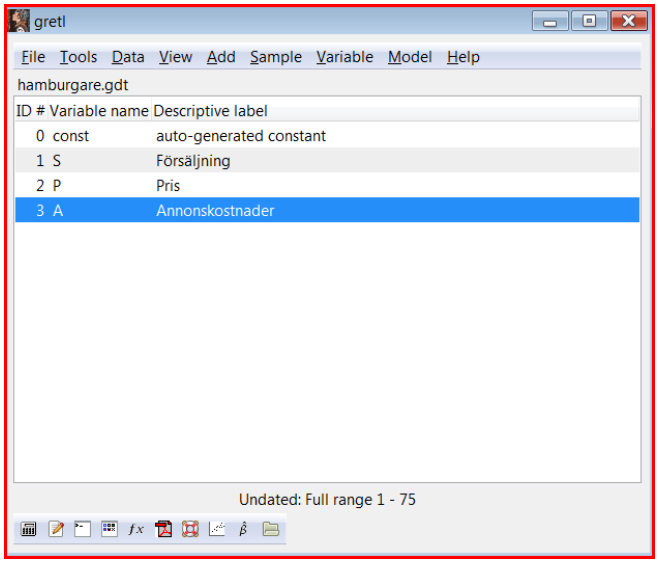
Parametrarna β_1, β_2,…, β_K mäter effekten av en enhets förändring i x_{ik} på det genomsnittliga värdet av y_i *om man håller alla andra variabler i ekvationen konstanta*.

Låt oss titta på ett exempel där vi har två oberoende variabler och en konstant.

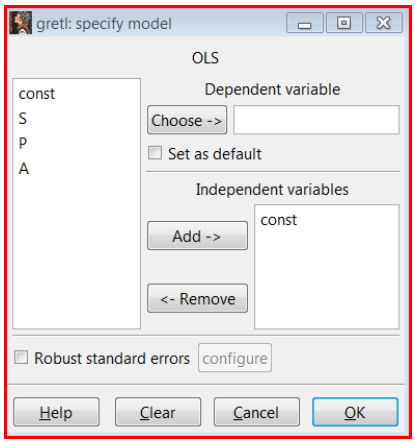
$$S_i = \beta_1 + \beta_2 P_i + \beta_3 A_i + e_i \quad = 1, 2, \ldots, N \qquad (5.2)$$

där S_i är månatlig försäljning mätt i 1000-tals kronor, P_i är priset på hamburgarna mätt i kronor och A_i är annonskostnader också mätt i kronor

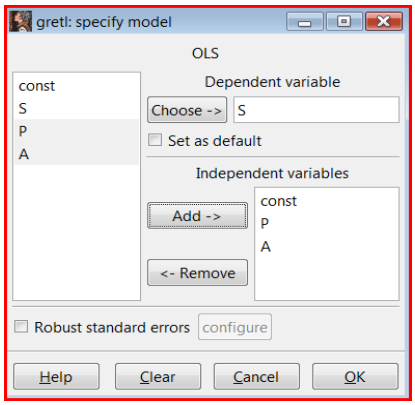
Börja med att ladda ner datafilen ”hamburgare.gdt”



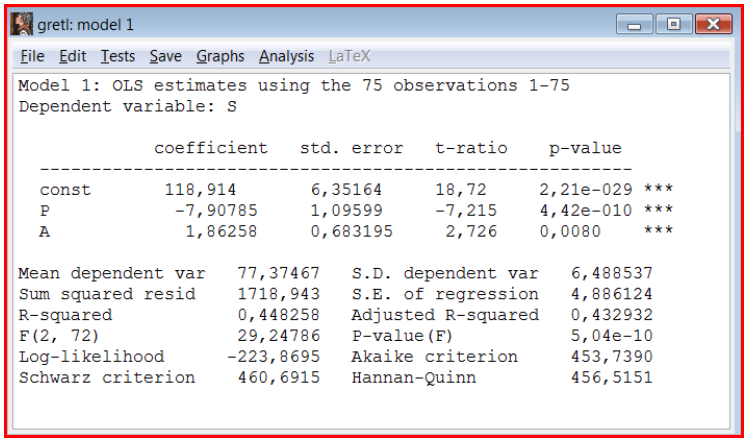
Välj sedan Model>Ordinary Least Squares.



Markera ”S” och klicka på knappen Choose så anger vi att S är den beroende variabeln. Markera sedan ”P” och ”A” och klicka sedan på knappen Add så läggs dessa två variabler in som oberoende variabler i skattningen:



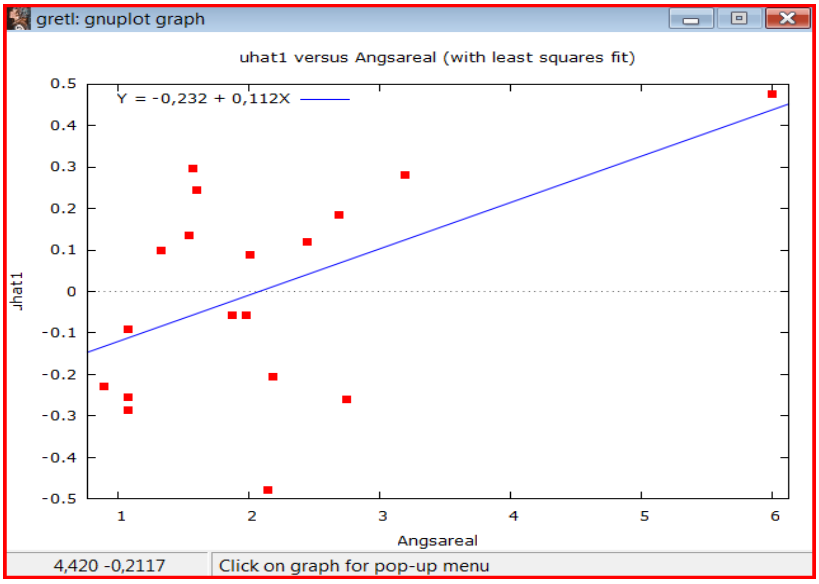
Klicka därefter på knappen OK. Du får då upp följande fönster:



Vi ser som förväntat att försäljningen tenderar att minska när priset ökar och att annonsering tenderar att öka försäljningen.

4.1 En multipel regressionsanalys av skördeutfall

I Eggeby & Söderbergs datamaterial från Luleå socken omkring 1850 finns förutom de tidigare angivna variablerna ytterligare en variabel i datafilen, Angsareal, och vi kan testa om det är så att ängsarealens storlek kan tänkas ha ett samband med kreatursinnehavet. Om vi gör ett spridningsdiagram med uhat1 - residualen från den tidigare enkla regressionsskattningen - och Angsareal (genom att markera dem i huvudfönstret och sedan högerklicka i någon av dem och i rullgardinsmenyn markera XY scatterplot och välja Angsareal som x-axelvariabel) får vi följande figur:



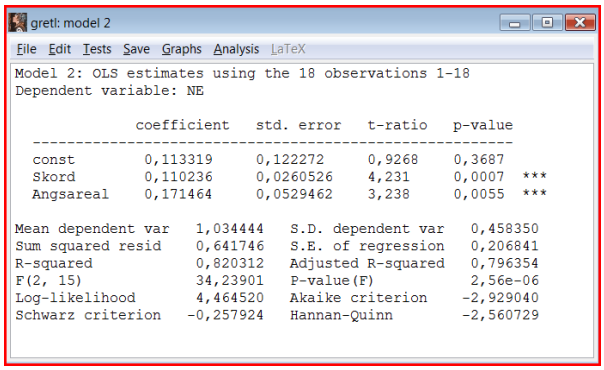
Här ser man tydligt att byar med liten ängsareal tenderar att ha negativa residualer och att de större byarna har större residualer. Detta tyder på att ängsareal skulle kunna vara en relevant oberoende variabel att ha med i regressionen. Vi får då en multivariat modell

$$\hat{y} = \beta_1 + \beta_2 x_2 + \beta_3 x_3$$

som vi kan skatta enklast i consolen med kommandot

ols NE const Skord Angsareal

Vi kan så klart också gå via Model i huvudfönstret och får då som output figuren nedan. Vi ser här att både Skord och Angsareal är ”statistiskt signifikanta”.
Modellen har nu ett r^2 på 0.82 och förefaller därför bättre ”förklara” storleken på nötkreatursenheter/cap bättre än modellen med bara en förklaringsvariabel.



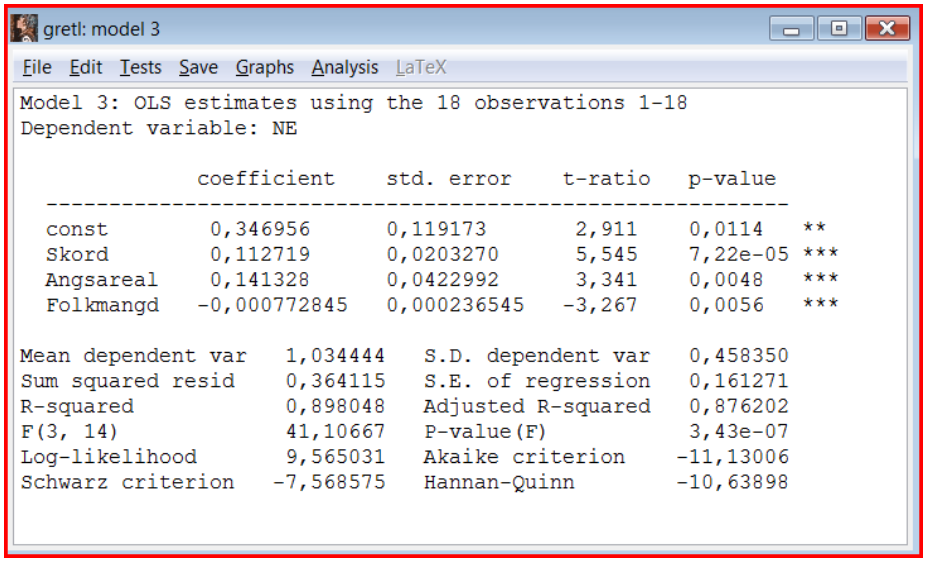
Modellen kan dock förbättras ytterligare genom att även inkludera variabeln Folkmangd. Vi skattar modellen

$$\hat{y} = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

med

ols NE const Skord Angsareal Folkmangd

eller som tidigare via Model och får som resultat:



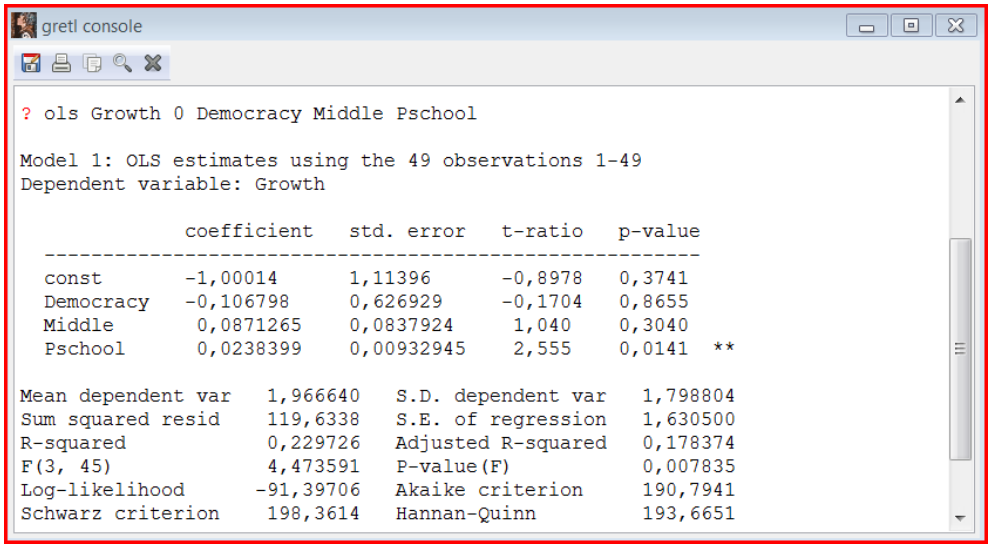
Denna modell med tre oberoende variabler ”förklarar” nästan 90 % av variationen i nötkreatursenheter/cap. Men titta också på parametervärdena! Är Folkmangd förutom att vara statistiskt relevant också historiskt relevant?

4.2 En multipel regressionsanalys av demokrati

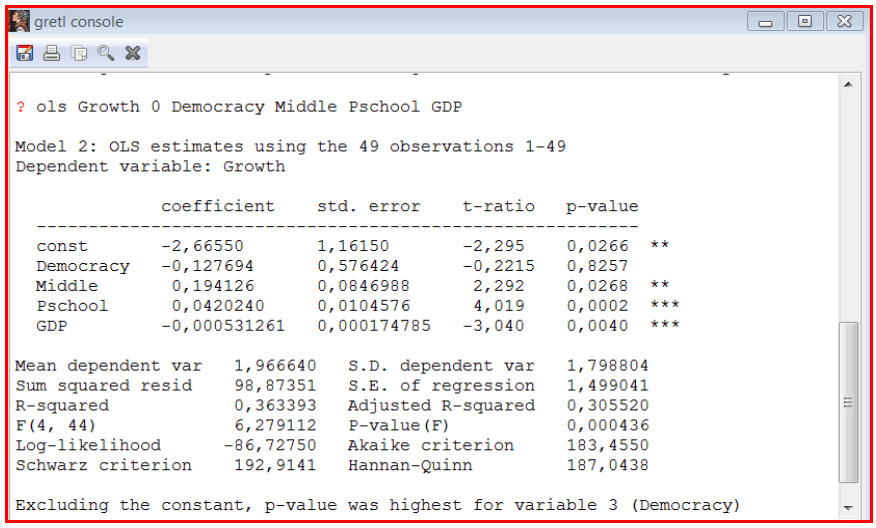
Låt oss än en gång titta på materialet i vårt dataset Tabell 5_1.gdt. Vi börjar med att testa hur demokrati, jämlikhet och bred skolgång påverkar tillväxten.
För att göra detta skattar vi modellen

ols Growth 0 Democracy Middle Pschool

och får då följande resultat:



Som vi ser är det bara Pschool som verkar ha en rimlig grad av signifikans. Låt oss lägga till BNP-variabeln GDP och se vad som händer:

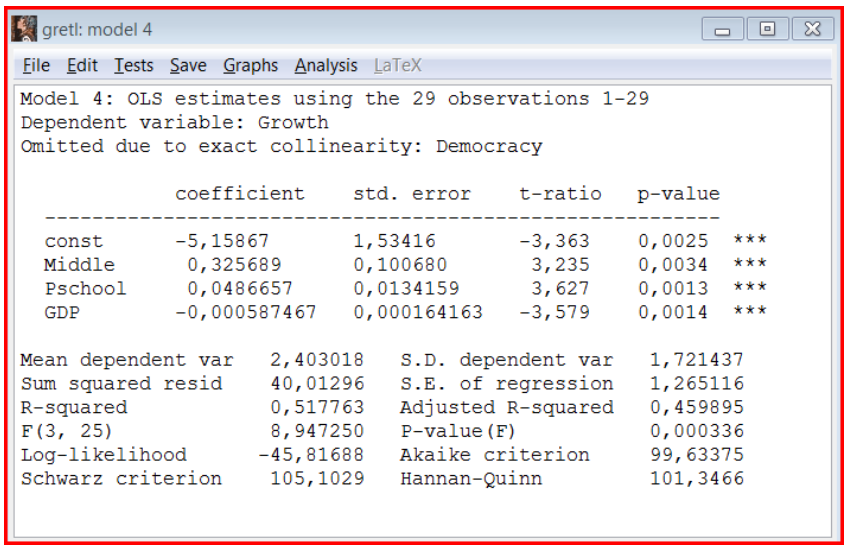


Nu blir alla variabler förutom Democracy ”signifikanta” och som vi ser har r^2 också ökat.

Kanske skulle man kunna misstänka att sambanden mellan variablerna ser olika ut i demokratier och diktaturer. För att kolla om så är fallet kan man göra separata skattningar för demokratier och diktaturer. För att göra detta går man tillväga på följande vis: I huvudfönstret väljer man Sample>Define, based on dummy och markerar Democracy i rullgardinsmenyn och klickar på knappen OK. I huvudfönstret ser vi då längst ner texten ”Undated: Full range = 49; current sample n = 29” vilket anger att vi nu bara tittar på de data som gäller för de 29 demokrati-länderna. Skatta nu åter modellen:

ols Growth 0 Middle Pschool GDP

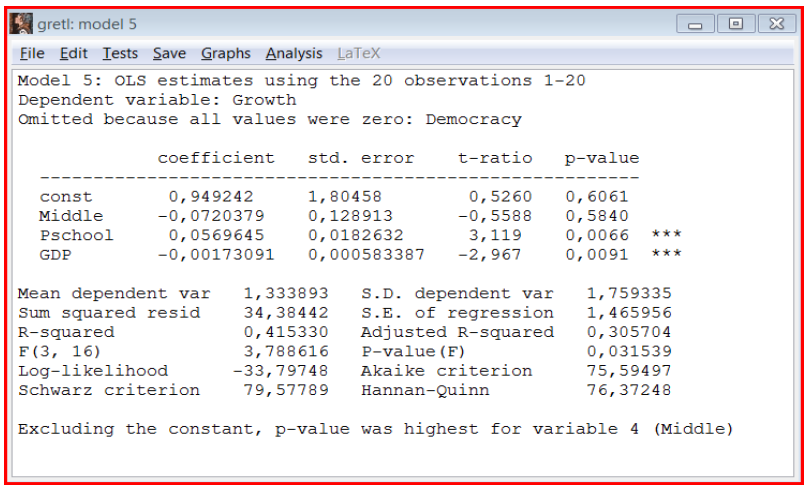
Vi får då följande resultat:



Som vi ser är alla variabler, inklusive jämlikhetsvariabeln Middle, ”signifikant”. Men hur ser det ut i diktaturer? För att kolla det går vi tillbaks till huvudfönstret och väljer Sample>Restore full range och sedan åter Sample>Define, based on dummy. Markera Dumnondem i rullgardinsmenyn och klickar på knappen OK. I huvudfönstret ser vi då längst ner texten ”Undated: Full range = 49; current sample n = 20” vilket anger att vi nu bara tittar på de data som gäller för de 29 diktatur-länderna. Skatta nu åter modellen:

ols Growth 0 Middle Pschool GDP

Vi får då följande resultat:



Det verkar som om jämn inkomstfördelning inte har samma effekter på tillväxten i demokratier och diktaturer.

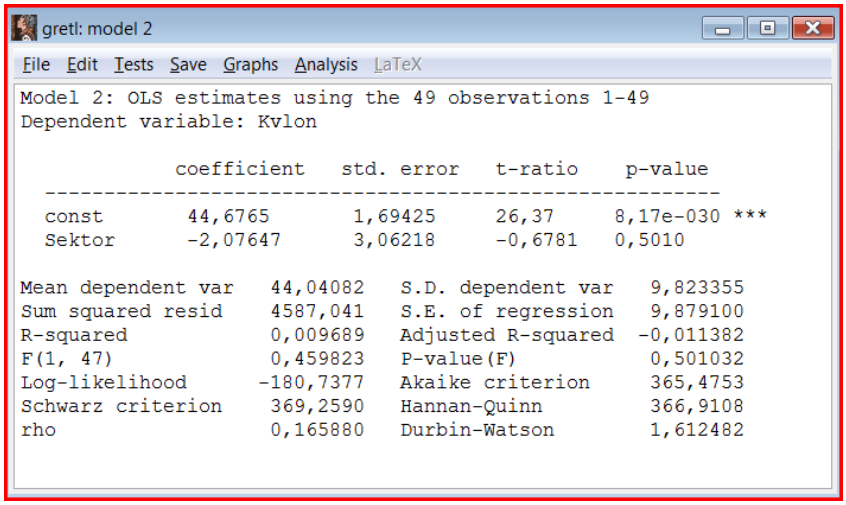
4.3 Ett exempel på variansanalys

Ett speciellt fall av regressionsanalys är så kallad variansanalys (ANOVA). **Gretl** saknar en del av de speciella funktionerna för traditionell variansanalys. Men det gör inget. Så länge de oberoende variablerna är diktotoma (det vill säg är av slaget att man kan sätta värdena 0 eller 1 på dem) går det nämligen lika bra att använda regressionsanalys, eftersom beräkningarna i en variansanalys sker på samma sätt som i regressionsanalysen. Skillnaden är huvudsakligen att man i regressionsanalysen även kan inkorporera oberoende variabler som också är kvantitativa (och inte bara kvalitativa variabler som i variansanalysen).

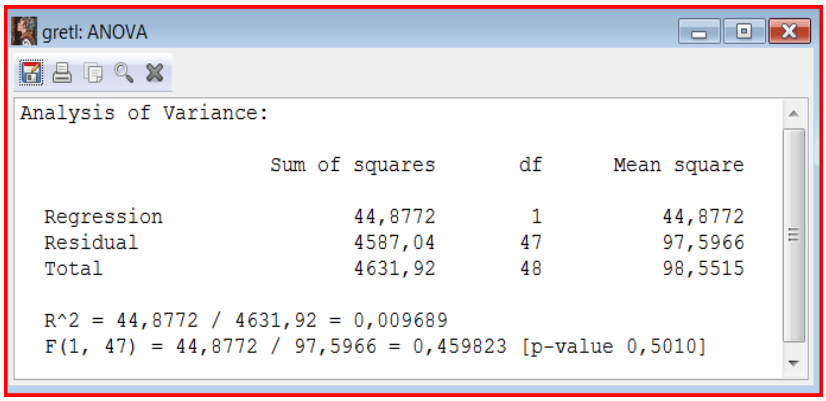
I Eggeby och Söderberg används data på kvinnolöner i den industriella revolutionen i England för att exemplifiera variansanalysmetoden. Låt oss titta på datan, som finns i Tabell 8_11 om kvinnolöner i jb och ind.gdt.
Kör en vanlig regression på kvinnolönerna i de två sektorerna:

ols Kvlon 0 Sektor

Vi får då



Vill vi se resultatet i form av en så kallad ANOVA-tabell, väljer vi fönstret Analysis > ANOVA och får följande uppställning (som motsvarar uppställningen på sidan 165 i Eggeby & Söderbergs bok):



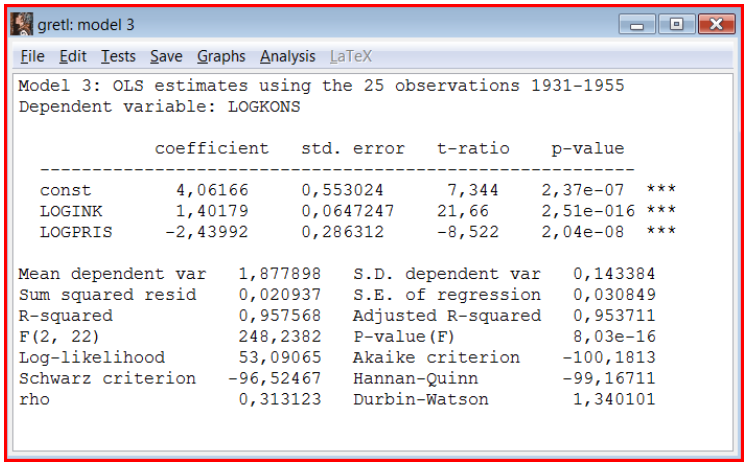
Som synes är värdet på r^2 väldigt lågt och det verkar som om det generellt sett inte hade någon betydelse om man arbetade inom jordbruk eller industri vad gäller relativförhållandet mellan kvinnors och mäns löner. Det förefaller som om det utifrån detta material inte går att hävda att den industriella revolutionen hade några tydliga effekter på kvinnor och mäns relativlöner.

4.4 Pris- och inkomstelasticiteter

Vi kan också använda multipel regressionsanalys om vi exempelvis vill analysera vilka effekter inkomst- och prisförändringar har på konsumtionen. I Eggeby & Söderberg finns ett exempel med data från 1931 till 1955 (index 1955=100). Ladda ner datafilen Restaurangexemplet.gdt. För att kunna beskriva de procentuella förändringarna på ett smidigt sätt har vi logaritmerat värdena på variablerna (se variablerna LOGINK, LOGKONS och LOGPRIS). Ta sedan

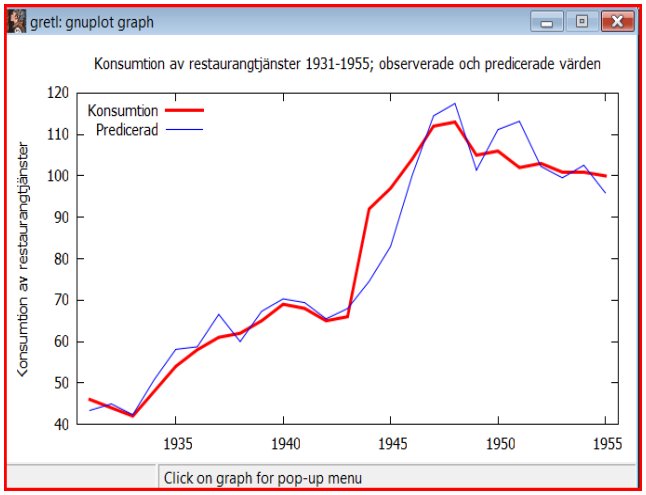
```
ols LOGKONS 0 LOGINK LOGPRIS
```

Resultatet blir

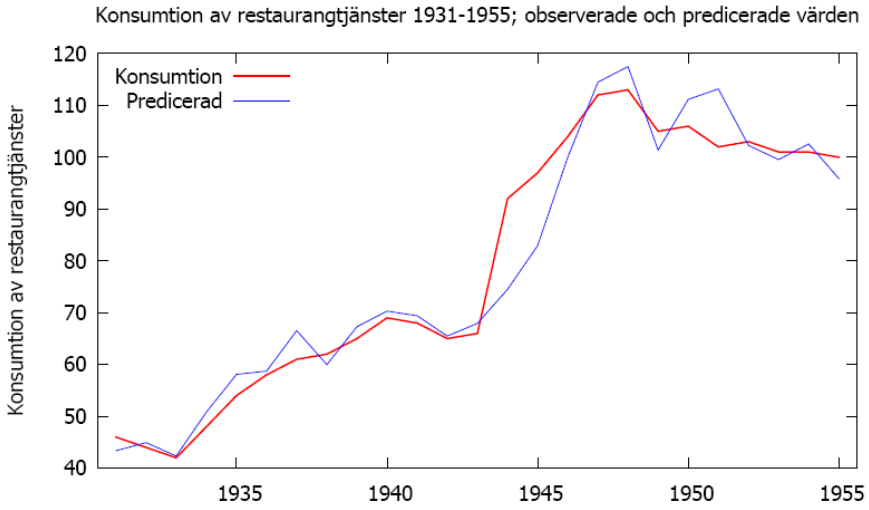


Värdet på inkomst-elasticiteten är 1.4, vilket innebär att om inkomsterna steg med 10% så ökade restaurangkonsumtionen med 14%. Priselasticiteten är 2.4, vilket innebär att en ökning av priserna med 10% medförde en minskad restaurangkonsumtion med 24%.

Med modellen kan vi också få fram predicerade värden (detta har vi redan gjort genom att välja Save > Fitted values och i rullgardinsmenyn välja att kalla variabeln för Predicerad). I huvudfönstret markerar vi variablerna Konsumtion och Predicerad och högerklickar i någon av dem och väljer Time series plot. Vi får då (efter lite editerande) följande figur:



Och om vi vill kunna använda grafen i någon publikation kan vi spara figuren som en pdf-fil:

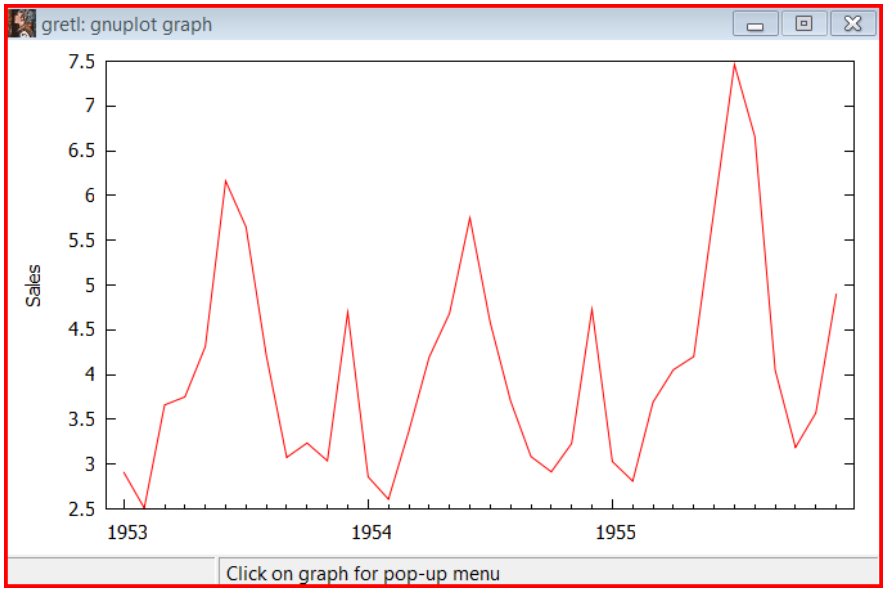


5 Tidsserieanalys

Gretls styrka som statistik- och ekonometriprogram ligger i dess mångsidiga möjligheter till olika slags analyser av tidsseriematerial. Tidsserieanalys är dock över lag svårare att genomföra än vad vi i övrigt tar upp i den här introduktionen och jag ska därför bara gå igenom några av de mer elementära metoderna.

5.1 Nedbrytning av en tidsserie i komponenter

I många tidsserier vill vi kunna bryta ner datan i olika komponenter. Om vi exempelvis har en tidsserie över försäljning av läskedrycker i Sverige 1953-1955 som ser ut så här (ladda ner Tabell 9_10.gdt och markera Sales, högerklicka och välj Time series plot):



Här verkar det som om det föreligger både trendmässigt en ökning och säsongsmässiga variationer. Trenden får vi genom att genomföra regressionen

```
ols Sales 0 tid
```

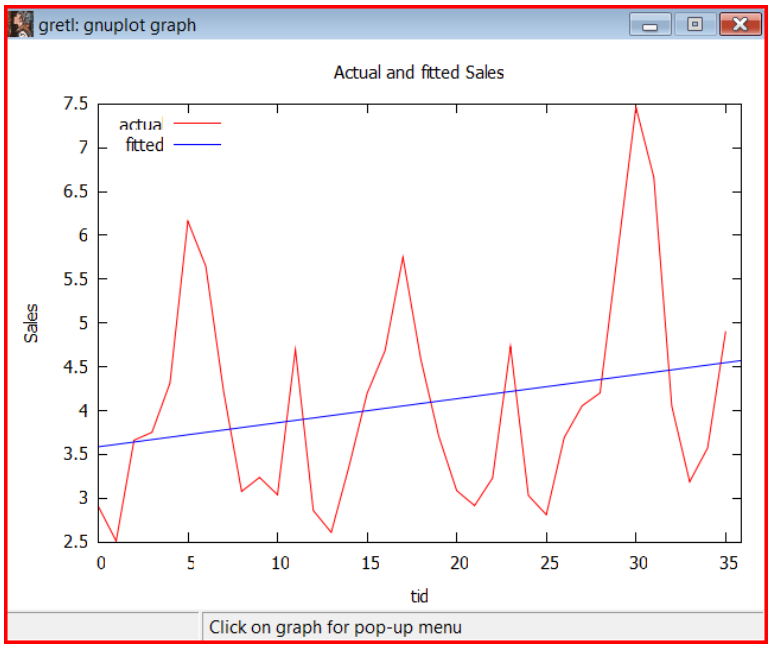
vilket ger outputen

gretl: model 2

File Edit Tests Save Graphs Analysis LaTeX

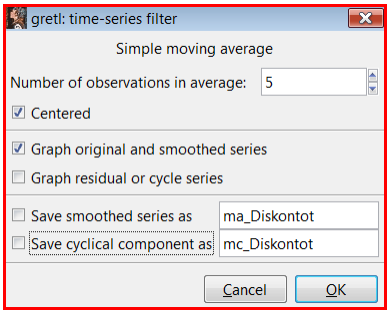
```
Model 2: OLS estimates using the 36 observations 1953:01-1955:12
Dependent variable: Sales

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       3,58633      0,388978      9,220    8,95e-011 ***
  tid         0,0274557    0,0191134     1,436    0,1600

Mean dependent var   4,066806   S.D. dependent var   1,209297
Sum squared resid    48,25539   S.E. of regression   1,191334
R-squared            0,057217   Adjusted R-squared   0,029488
F(1, 34)             2,063431   P-value(F)           0,160009
Log-likelihood      -56,35558   Akaike criterion     116,7112
Schwarz criterion    119,8782   Hannan-Quinn         117,8165
rho                  0,561832   Durbin-Watson        0,867116
```

Regressionslinjen kan läggas in i figuren genom att i modellfönstret välja Graphs > Fitted, actual plot > Against tid och i gnuplotgrafen högerklicka, markera Edit, trycka på knappen lines och i rullgardinen vid sidan om type välja ”lines” och sedan OK. Då får vi

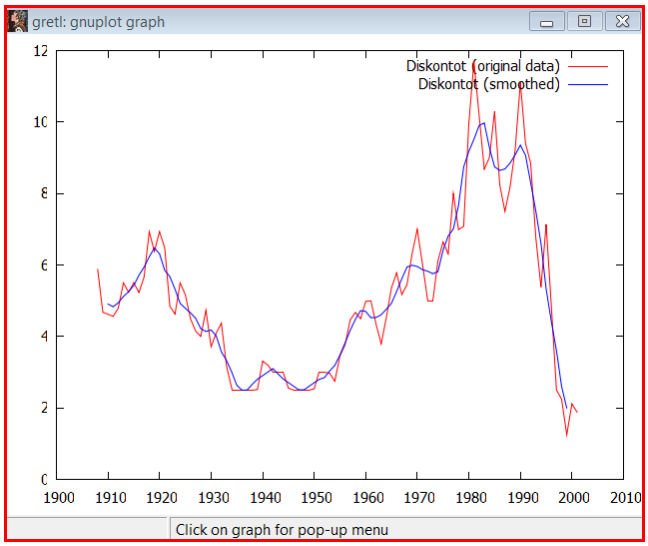


Om vi sparar trendvärdena (vilket vi gjort som Trendvarde i datafilen) kan man sedan som i Eggeby & Söderberg beräkna säsongsvariationen genom att ta medelvärden för de månadsvisa avvikelserna (resultatet finns i variabeln Seasonal). Men här visar **gretl** sin verkliga styrka. I stället för att genomföra dessa ganska tidskrävande operationer kan man i stället välja att markera Sales i huvudfönstret och välja Variable>Tramo-analysis eller X-12-ARIMA-analysis och få fram en mer avancerad dekomponering.

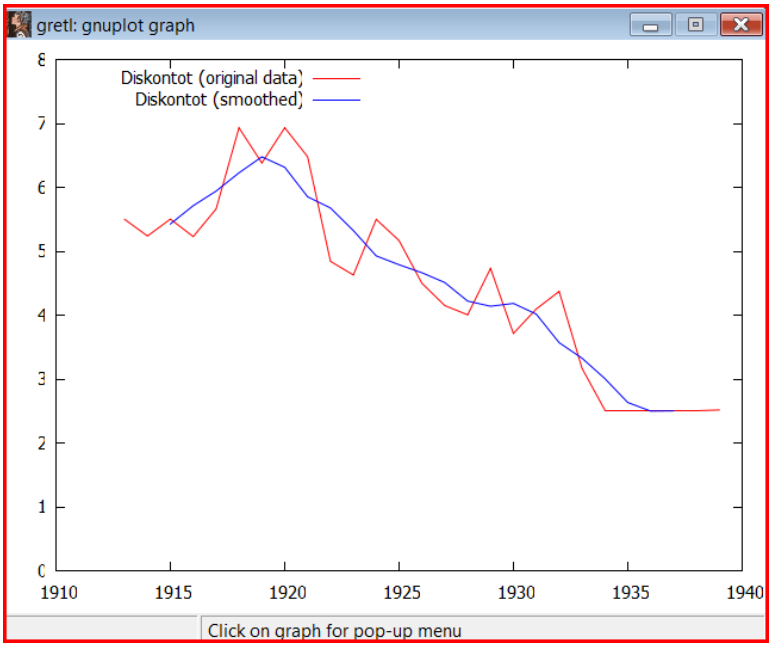
En alternativ form av trendbestämning som ofta använts av ekonomihistoriker är att ta löpande medeltal. Låt oss titta på variabeln Diskontot i datafilen Tabell 9_5 Diskontot 1908_2001.gdt. Markera variabeln, välj Variable Filter>Simple moving average och markera så här:



Tryck sedan på OK och vi får följande graf:



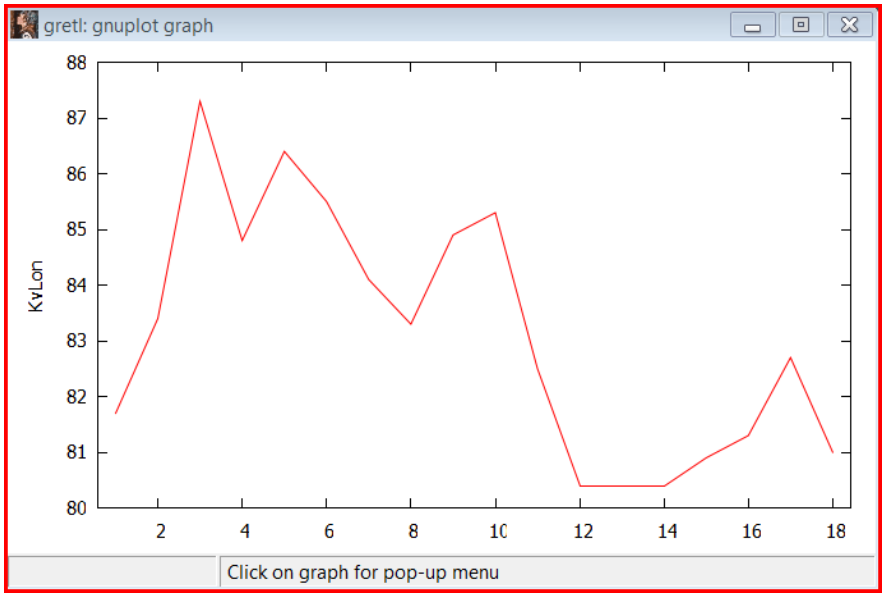
Sätter vi för jämförelse med Figur 9.5 i boken samplet till 1913-1939 får vi (efter lite editerande):



5.2 Trendbestämning med regressionsanalys

Regressionsanalys kan som vi sett användas för att testa om det över tiden finns en trend i en data serie. Tiden används då som den oberoende variabeln och det gäller att utifrån denna se om den beroende variabeln uppvisar en trend över tiden eller ej.

Låt oss börja med att titta på datafilen Tabell 9_13.gdt, som innehåller data över kvinnors procentuella lön i förhållande till männens. Låt oss först titta på tidsserieploten av KvLon:



Här kan det vara svårt att tydligt se en trend och vi kör därför regressionen

ols KvLon 0 Tidstrend

för att se om det finns en trend. Resultatet blir:

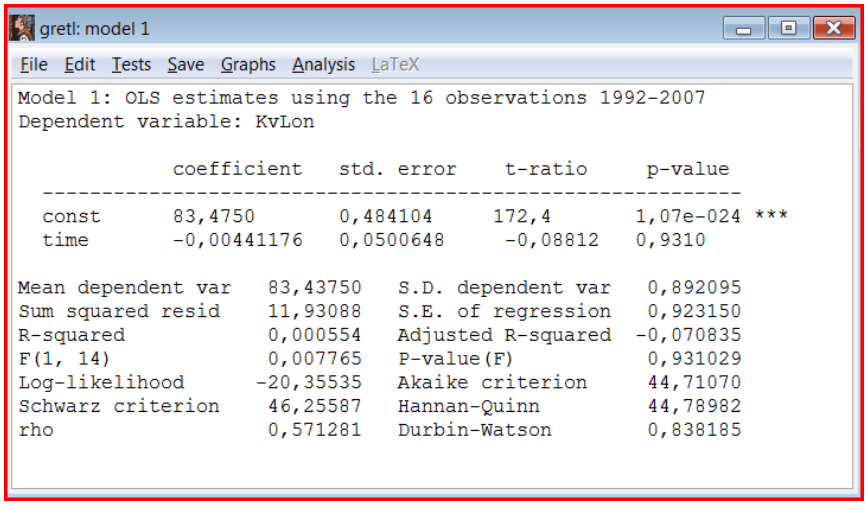
gretl: model 4

File Edit Tests Save Graphs Analysis LaTeX

```
Model 4: OLS estimates using the 18 observations 1-18
Dependent variable: KvLon

              coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const        85,8206      1,06923      80,26     2,79e-022 ***
  Tidstrend    -0,219329    0,0792931    -2,766    0,0138    **

Mean dependent var   83,12778   S.D. dependent var   2,212479
Sum squared resid    56,29593   S.E. of regression   1,875765
R-squared            0,323497   Adjusted R-squared   0,281216
F(1, 16)             7,651050   P-value(F)           0,013772
Log-likelihood      -35,80315   Akaike criterion     75,60629
Schwarz criterion    77,38704   Hannan-Quinn         75,85184
rho                  0,378075   Durbin-Watson        0,973287
```

Här ser vi att regressionskoefficienten för trenden är -0.22 (och vi ser på p-värdet att sannolikheten för att värdet skulle vara noll är mindre än två procent). Det innebär att över hela perioden 1975-1995 har kvinnornas relativlöner i snitt fallit med 0.22 procent per år.

Hur ser det ut om vi gör en liknande analys för en något senare period? Låt oss se på en dataserie från SCB som finns sparad som KvLon SCB 1992_2007 ej standardvägda.gdt Här får vi följande resultat:



Här verkar det som att det inte längre finns någon trend eftersom värdet på tidskoefficienten dels är väldigt litet och dels insignifikant (p-värde = 0.93). Här kan vi inte längre, för den aktuella perioden, se något som indikerar att det är sannolikt att kvinnornas medianlöner har fallit relativt männens.

6. Några grundläggande sannolikhetsbegrepp

Eftersom de värden som exempelvis ekonomiska variabler har inte är kända innan de observeras säger vi att de är *slumpmässiga* eller *stokastiska*. Sannolikhetsteorin ger oss möjligheter att analysera denna typ av variabler där det alltid ingår ett mått av osäkerhet om vilka värden variablerna tar. Varje gång vi observerar utfallet av en slumpvariabel får vi en observation. När vi väl observerat variabelvärdet är det inte längre slumpmässigt. Det är alltså skillnad på variabler vars värde ännu inte observerats (slumpvariabler) och de vars värde redan observerats (observationer). En sannolikhetsfördelning är en matematisk beskrivning av de olika möjliga värden som slumpvariabeln kan ta. Slumpvariabeln ”antal flickor i en slumpmässigt vald trebarnsfamilj” kan anta värdena 0, 1, 2 eller 3 (dessa värden är variabelns utfallsrum). Sannolikhetsfördelningen talar om vilken relativ frekvens de olika utfallen har.

Vanligtvis brukar man fokusera på några centrala numeriska karakteristika som fördelningarna har. Dessa kallas *parametrar* och kan exempelvis vara *väntevärde* och *varians*. Som statistiker eller ekonometriker försöker man skatta dessa parametrar genom att använda den information som finns tillgänglig.

Om vi tar exemplet med flickorna kan m an fråga sig vad sannolikhetsfördelningen för X = antalet flickor i en slumpmässigt vald trebarnsfamilj är. Detta kan vi egentligen bara svara på om vi känner populationen. Oftast gör vi inte detta, men kan ändå försöka få en bild av fördelning genom att använda en *sannolikhetsmodell*.

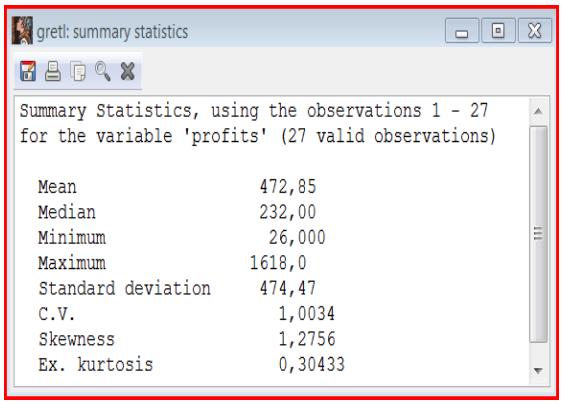
Om vi antar att sannolikheten för att få en flicka är 50% och inte påverkas av eventuella tidigare syskons kön skulle vi kunna beskriva väntevärde som

$$\mu = E[X] = \sum \mathrm{xp(x)} =$$

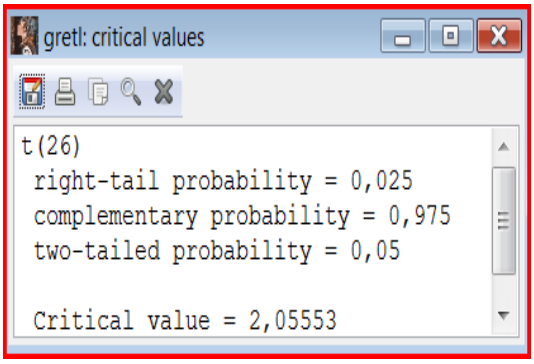
$$0 \cdot 0.125 + 1 \cdot 0.375 + 2 \cdot 0.375 + 3 \cdot 0.125 = 1.5$$

Variabeln X måste anta ett av värdena 0, 1, 2 eller 3 och därför blir summan av sannolikheterna lika med ett.

Här är μ alltså en parameter vars värde vi kan erhålla *om* vi känner fördelningsfunktionen för X. Men eftersom vi inte gör det får vi försöka skatta den med hjälp av ett *urval* (stickprov). Och med hjälp av **gretl** finner vi lätt medelvärden, varians, standardavvikelse, kovarians och andra stickprovsvärden. Det enklaste sättet är att helt enkelt markera en variabel (profits) i en datafil (data 3-10.gdt) och välja View > Summary statistics från rullgardinsmenyn.



Låt oss anta att vi vill skatta ett konfidensintervall för medelvärdet på variabeln profits. Eftersom den ”sanna” variansen, σ^2, inte är känd, får vi använda en t-fördelning och skatta den. Eftersom N = 27 blir frihetsgraden (df) N – 1 = 26. Välj Tools > pvalue finder > t och skriv in 26 vid ”df” och 0.025 vid ”value”. Du får då

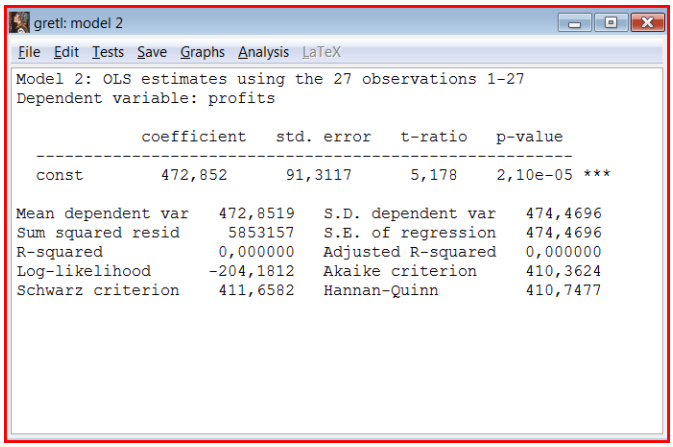


Det kritiska värdet för t_{26} fördelningen är som synes 2.055 och med hjälp av ett kommandoskript kan du få fram intervallet:

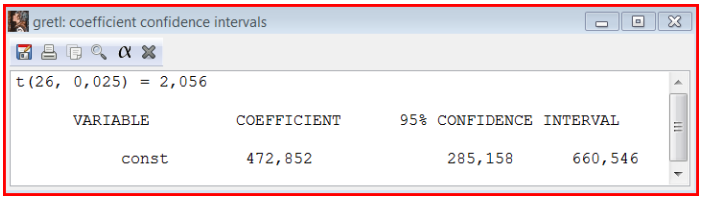
```
genr s2hat=sum((profits-mean(profits))^2)/($nobs-1)

genr varYbar=s2hat=s2hat/$nobs

genr sdYbar=sqrt(varYbar)

genr lobo=mean(profits)-2.055*sdYbar

genr hibo=mean(profits)+2.055*sdYbar
```

Intervallet är [285.158, 660.546] och du kan med 95% säkerhet vara ”säker” på att medelvärdet för profits ligger i detta intervall.

Vill du göra det ännu lättare för dig kan du i huvudfönstret välja Model > Ordinary Least Squares och markera profits och trycka på Choose och sedan OK. Då får du upp följande fönster:



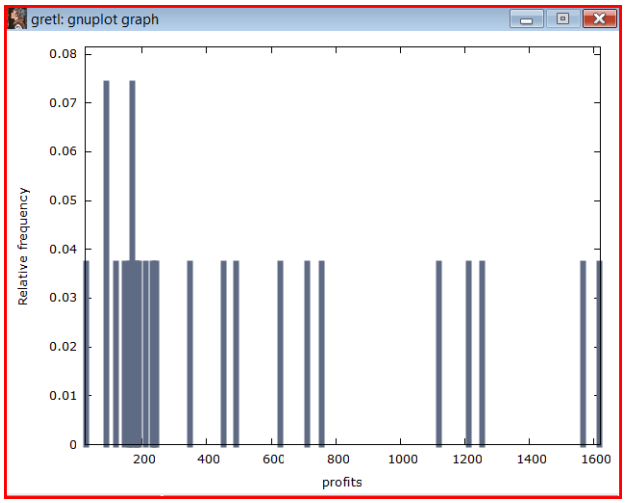
Här väljer du Analysis > Confidence intervals for coefficients och får intervallet i fönstret. Lättare än så här kan det inte bli!



7. Diagram, tabeller, centralvärde och spridningsmått

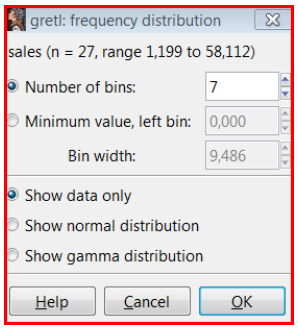
7.1 Stapeldiagram

Den enklaste formen av diagram är de som presenterar en enda variabel. I ett stapeldiagram avsätts frekvensen av variabeln på y-axeln och variabelvärdena på x-axeln. I **gretl** kan det se ut så här (efter att du i huvudfönstret markerat variabeln profits, högerklickat och valt Frequency plot i rullgardinsmenyn i data 3-10.gdt):

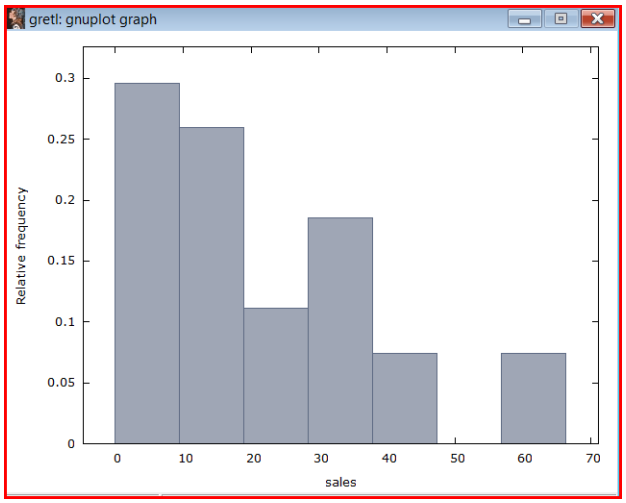


7.2 Histogram

Om variabeln är kontinuerlig använder man sig av ett histogram. Markera variabeln sales, välj Variable > Frequency plot och följande fönster dyker upp

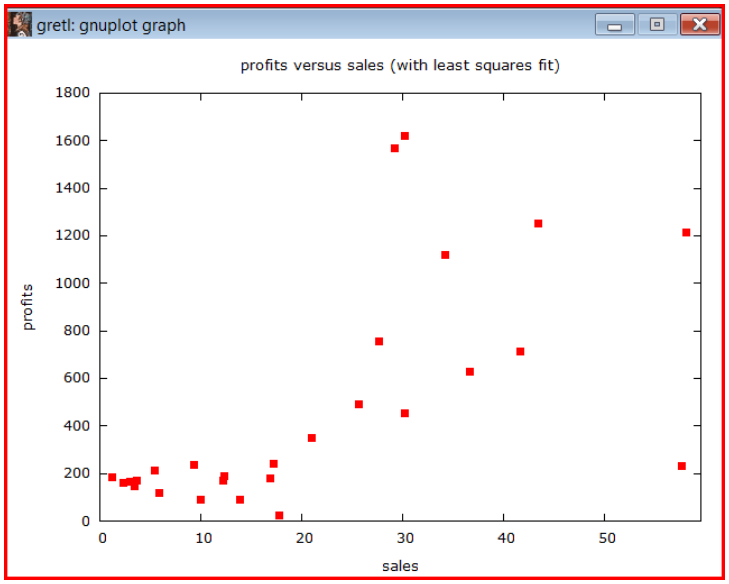


Vill du inte själv klassindela materialet (ett förslag på antal klasser, och indirekt klassbredden, anges automatiskt i rutan till höger om ”Number of bins”) trycker du bara på OK och i diagrammet som dyker upp klickar du på Edit, sedan på Lines och vid ”type” väljer du i rullgardinen att markera boxes, klickar på OK och får då följande histogram:



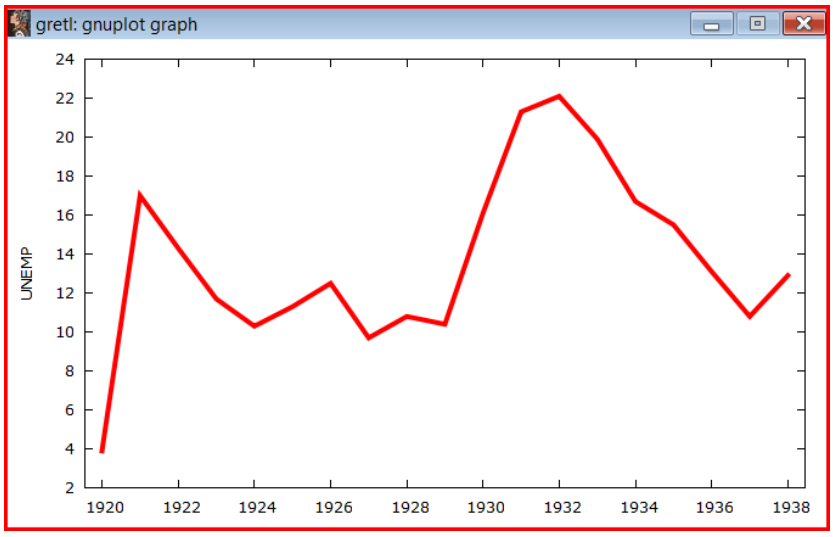
7.3 Sambandsdiagram

Vill vi se om det finns något samband mellan två variabler kan man som vi tidigare visat välja att göra ett prickdiagram (sambandsdiagram, på engelska xy-plot):



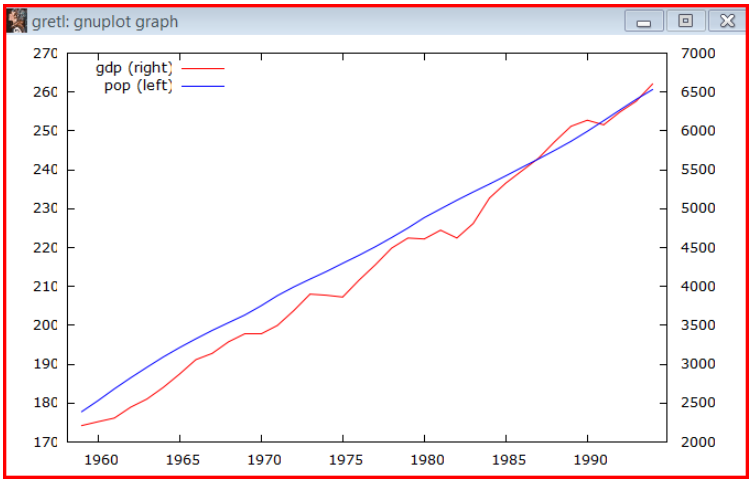
7.4 Tidsseriediagram

Vill vi göra ett tidsseriediagram (kurvdiagram) kan vi till exempel ta en datafil med uppgifter om arbetslösheten i England 1920-1938 (datafil UNEMP IN GB.gdt i HISTLIC-mappen). Markera UNEMP i huvudfönstret, högerklicka och välj Time series plot i rullgardinen och du får följande tidsseriediagram:

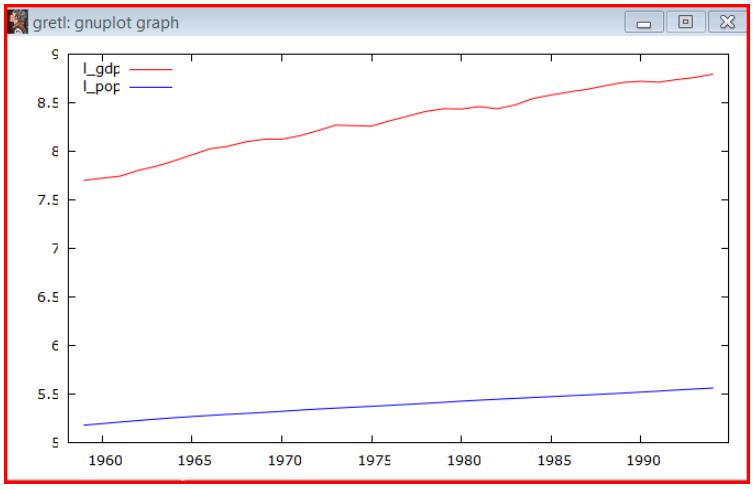


Här har vi ett diagram som lämpar sig väl om vi vill illustrera absoluta förändringar. Vill vi däremot fokusera på relativa förändringar är det mer ändamålsenligt att använda ett *semilogaritmiskt diagram*.

Ladda ner datafilen data3_15-gdt i Ramanathan-mappen och gör ett tidsseriediagram på variablerna gdp och pop:



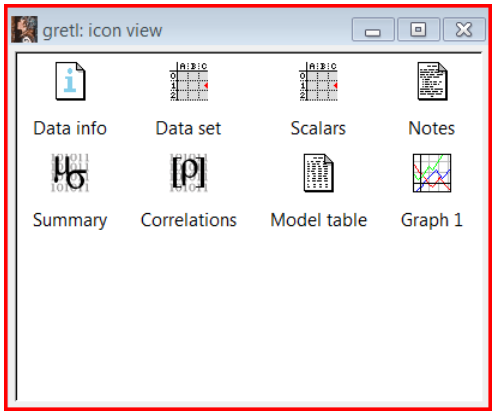
Diagrammet visar BNP och befolkningens utveckling i USA mellan åren 1959 och 1994. Om vi logaritmerar variablerna får vi följande tidsseriediagram:



Det semilogaritmiska diagrammet visar på den procentuella förändringen av variablerna från år till år.

7.5 Andra typer av diagram

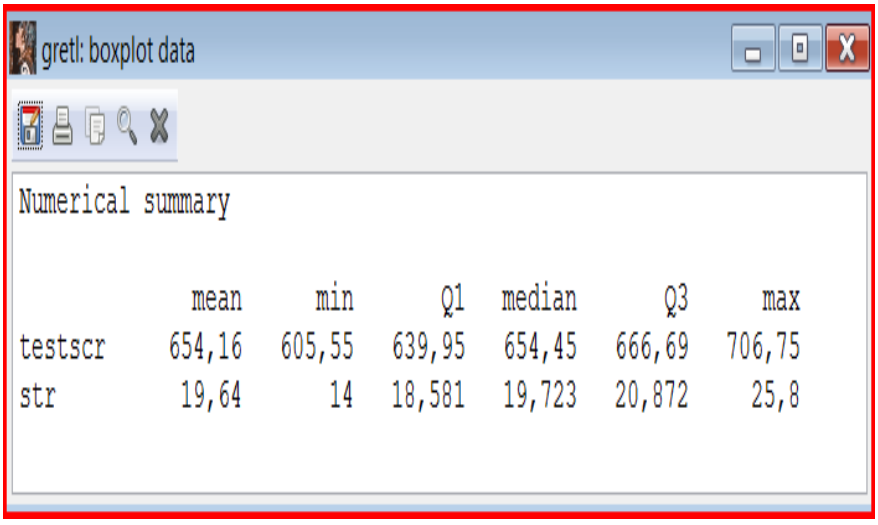
Du kan göra i princip vilken typ av diagram som helst i **gretl**. Men ska du göra andra än de jag beskrivit här krävs det lite extra arbete. **Gretl** utnyttjar ett program som heter **gnuplot** och det är via det som du kan utforma diagram och tabeller helt som du själv vill. Låt oss säga att du vill göra en mer avancerad variant av tidsseriediagrammet ovan. Du gör då på följande vis: Högerklicka i diagrammet, välj Save to session as icon. Tryck sedan på knappen session icon view i huvudfönstret (fjärde från vänster längst ner). Du får upp följande fönster. I denna högerklickar du på Graph1 och väljer Edit plot commands i rullgardinen. Hur du sedan kan gå tillväga kan du läsa mer om i kapitel 7 i Gretlguiden.



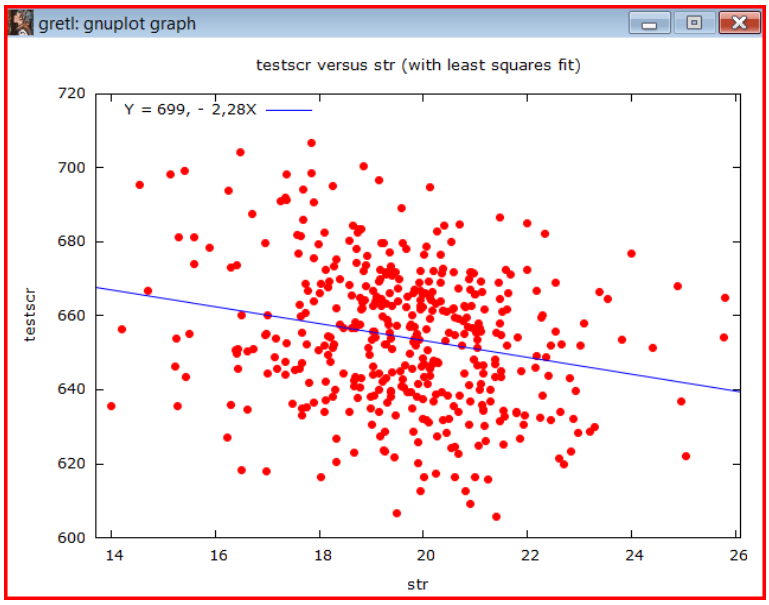
7.6 Ett skolexempel

Låt oss ta ett exempel med anknytning till skolans värld. Under lång tid har man diskuterat om det föreligger något samband mellan klasstorlek och elevernas prestation i skolan. I Kalifornien genomfördes år 1998 en undersökning i 420 skoldistrikt om det förelåg ett sådant samband.

För att först få ett överskådligt grepp om datan väljer vi i datafilen caschool.gdt - som vi får fram genom att i huvudfönstret välja File > Open data Sample file > Stock&-Watson (eller öppnar ifrån HISTLIC-mappen) - att klicka View > Graph specified variables > Boxplots och skriver in de två variablerna str (som anger hur många elever det går per lärare) och testscr (som anger genomsnittligt testresultat). Därefter klickar vi på OK och högerklickar sedan i grafen som dyker upp och väljer Numerical summary. Vi får då följande resultat:

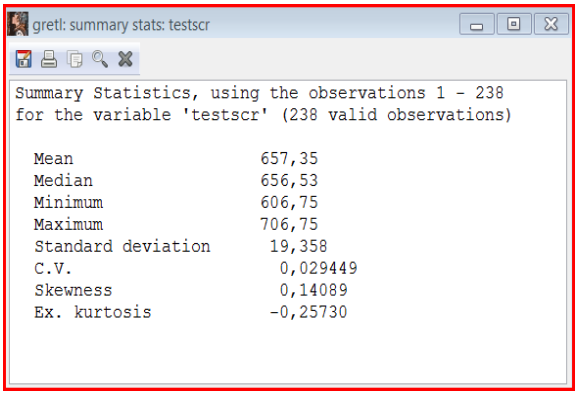


För att kunna få svar på frågan måste vi dock gå vidare och få fram ett sambandsdiagram. Markera de två variablerna i huvudfönstret, högerklicka på någon av dem och välj XY scatterplot. Som x-axelvariabel väljs str. Klicka sedan på OK och följande diagram dyker upp:

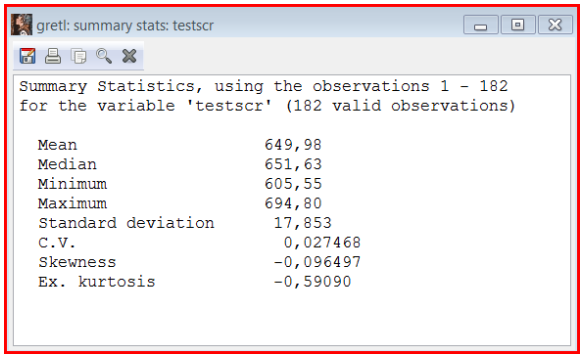


Här har automatiskt en OLS- skattning lagts in och vi ser att den negativa lutningen på regressionslinjen är -2.28 (om vi vill kan vi också markera de två variablerna i huvudfönstret, högerklicka och välj Correlation matrix och får då ett värde på -0.23). Vi finner alltså med några få knapptryckningar i **gretl** att det preliminära svaret på vår fråga är ja.

En preliminär analys av datan visar att om vi väljer distrikt med litet värde på str (< 20) och distrikt med stort värde (> 20) så uppvisar de stora skillnader i testresultat. Enklast får vi fram detta genom att välja Sample > Restrict, based on criterion och sedan i tur och ordning välja str < 20 och str ≥ 20, markera testcr och jämföra värdena vi får på Descriptive statistics. Vi får då

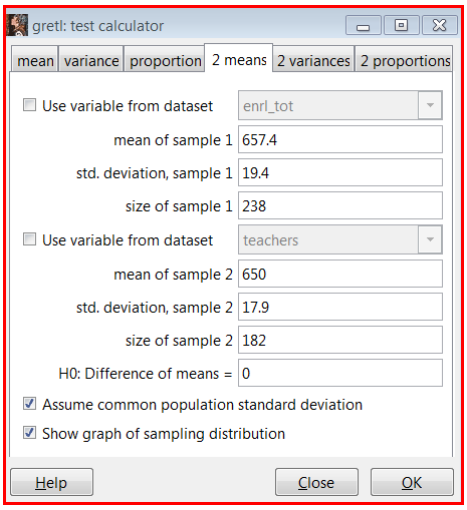


respektive

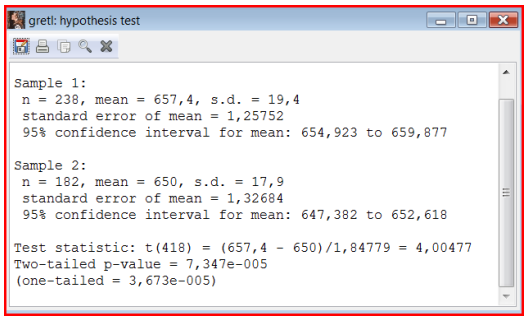


För den förra gruppen är medelvärdet på testresultatet 657.4 och för den senare 650.

Om skillnaden är statistiskt ”signifikant” tar vi lättast reda på via Tools test statistic calculator > 2 means och fyller i övriga fält som behövs:



Tryck på OK och resultatet blir



t-värdet är 4.005 och alltså större än 1.96, vilket innebär att vi på 5% nivån kan förkasta hypotesen att det inte skulle föreligga någon skillnad mellan medelvärdena (ekvivalent ser vi också att de 95% konfidensintervall för medelvärdena inte överlappar).

Står vi inför ett policyproblem - att exempelvis bestämma hur stora effekterna på testresultaten skulle bli om man minskade på antalet elever per lärare – är vi intresserade av relationen Δtestcore/Δstr. Vi gör en enkel regression

ols testscr 0 str

och får

gretl: model 3

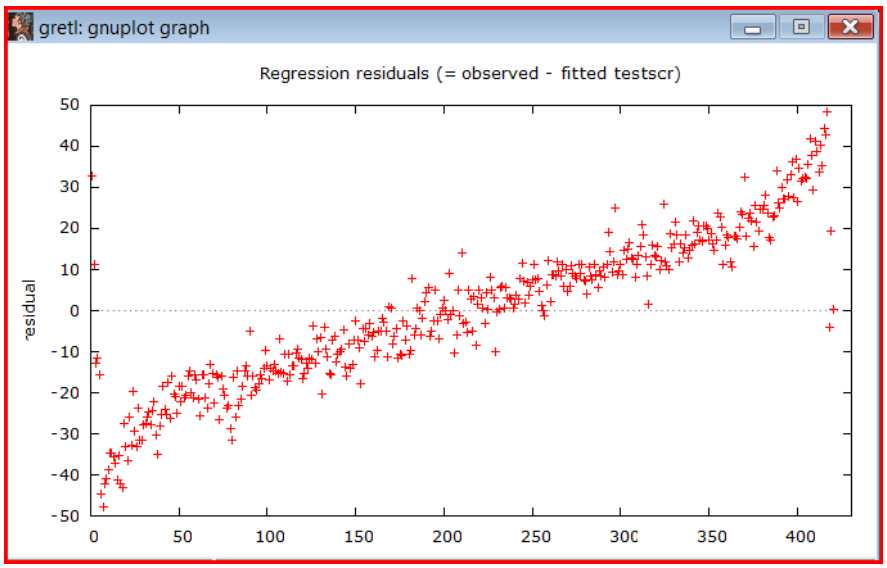
File Edit Tests Save Graphs Analysis LaTeX

```
Model 3: OLS estimates using the 420 observations 1-420
Dependent variable: testscr

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       698,933       9,46749     73,82     6,57e-242 ***
  str          -2,27981     0,479826    -4,751    2,78e-06  ***

Mean dependent var   654,1565   S.D. dependent var   19,05335
Sum squared resid    144315,5   S.E. of regression   18,58097
R-squared            0,051240   Adjusted R-squared   0,048970
F(1, 418)            22,57511   P-value(F)           2,78e-06
Log-likelihood      -1822,250   Akaike criterion     3648,499
Schwarz criterion    3656,580   Hannan-Quinn         3651,693
```

Om alla andra faktorer konstanthålls innebär detta (fundera kring extern validitet här) att en minskning av antal elever per lärare med en enhet skulle höja testresultatet med 2.28 enhet. Δtestcore/Δstr = -2.28.

Notera också att även om förändringen av str endast ”förklarar” en väldigt liten del av förändringen i testcr (r^2 = 0.05) så är beroendet starkt signifikant (titta på p-värdet och ”S.E. of regression”, som är √Sum squared resid/(n-2)).

Om vi tar skoldistrikte Antelope (som är observation 242 i vårt datamaterial), så har det str = 19.33 och testcr = 657.8. Utifrån vår regressionsmodell skulle vi skatta distriktets testresultat till 654.8 (698.9 – 2.28·19.33). Residualen skulle bli 3.0 (657.8 – 654.8). Vi kan göra samma beräkningar för alla skoldistrikten och se residualerna i ett diagram (klicka i modellfönstret på Graphs > Residual plot och välj By observation number i rullgardinsmenyn):



Över lag måste man nog konstatera att vi inte kan säga så mycket om testresultaten enbart utifrån antal elever per lärare. Fler faktorer spelar in och vi ska definitivt vara mycket försiktiga med att dra några som helst kausala slutsatser. Det kan mycket väl vara så att

det finns ett flertal oidentifierade variabler som påverkar den oberoende variabeln, vilket kan innebära att deras effekter så att säga samlas upp av variabeln str. Sambandet blir därigenom missvisande.
[Man brukar argumentera för att dessa oidentifierade variabler inkluderas i residualen, ε. Om det är så kan man säga att OLS-regressionen konsistent skattar de konditionala förväntade värdena på testresultaten givet bland annat värdet på variabeln str. Men regressionen skattar *inte* konsistent den *kausala* effekten av str. En OLS-skattning av str-parametern β_2 återspeglar skillnaden i förväntade testresultat mellan vilka som helst två skoldistrikt som *bara* skiljer sig åt vad gäller str. Den mäter inte den förväntade skillnaden i testresultat om ett skoldistrikt väljer att ändra sitt str. Anledningen är att när vi tolkar modellen som en konditional förväntan (E(testcr|str)) så antar vi inte att de andra, oidentifierade, variablerna är konstanta i de olika skoldistrikten, medan i den kausala tolkningen de antas vara konstanta. När vi tolkar modellen i termer av konditionala förväntningar så refererar *ceteris paribus*-villkoret endast till variabeln str, medan vid en kausal tolkning även de oidentifierade variablerna som anges av residualtermen inbwegrips i *ceteris paribus*-villkoret. Vi måste således vara ytterst försiktiga när vi tolkar våra resultat i kausala termer! Till exempel är det inte ovanligt att den förväntade lönen för gifta och ogifta arbetstagare varierar även om det inte är speciellt troligt att giftermålet i sig orsakar löneskillnaden. Snarare är det väl så att den maritala statusen fungerar som en ”proxy” för en rad andra karakteristika som påverkar en persons lön. Ska vi tolka en löneregressions koefficienter som kausala måste *alla* andra faktorer - och inte bara de observerbara variabler vi råkar inkludera i modellen - inkluderas under *ceteris paribus* antagandet. Huruvida detta är rimligt kan bara avgöras från fall till fall. Och tyvärr hjälper oss statistiska test genomgående väldigt lite i denna fråga.]

Om vi utökar modellen genom att också ta hänsyn till hur många procent av eleverna i distriktet som lär sig engelska som andraspråk, variabeln el_pct, får vi en multipel regressionsmodell

```
ols testscr 0 str el_pct
```

som resulterar i

gretl: model 1

File Edit Tests Save Graphs Analysis LaTeX

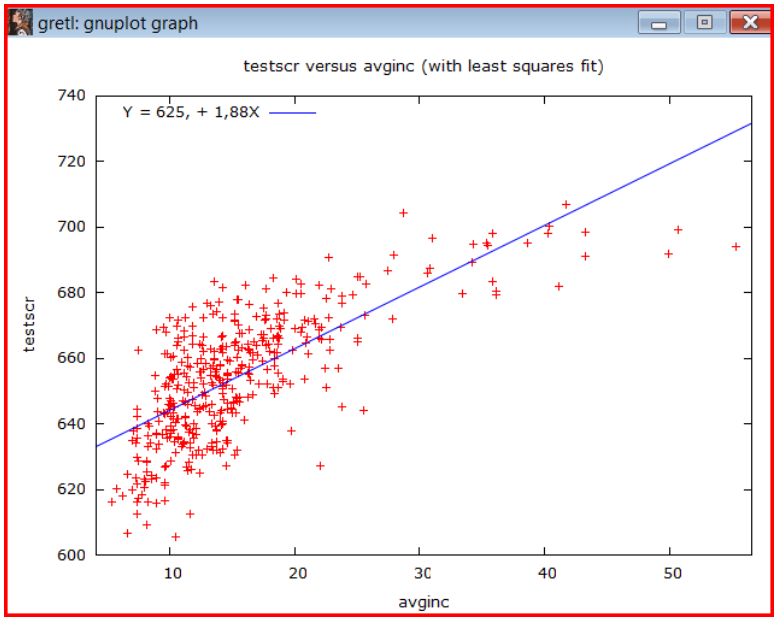
```
Model 1: OLS estimates using the 420 observations 1-420
Dependent variable: testscr

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      686,032       7,41131       92,57    3,87e-280 ***
  str         -1,10130     0,380278      -2,896   0,0040    ***
  el_pct      -0,649777    0,0393425    -16,52    1,66e-047 ***

Mean dependent var   654,1565   S.D. dependent var   19,05335
Sum squared resid    87245,29   S.E. of regression   14,46448
R-squared            0,426431   Adjusted R-squared   0,423680
F(2, 417)            155,0136   P-value(F)           4,62e-51
Log-likelihood      -1716,561   Akaike criterion     3439,123
Schwarz criterion    3451,243   Hannan-Quinn         3443,913
```

Koefficienten på str säger oss nu att om vi ökar denna variabel med en enhet så kommer - *om man håller procenten av elever som lär sig engelska som andraspråk konstant* – så kommer testresultaten att minska med 1.1 enheter.

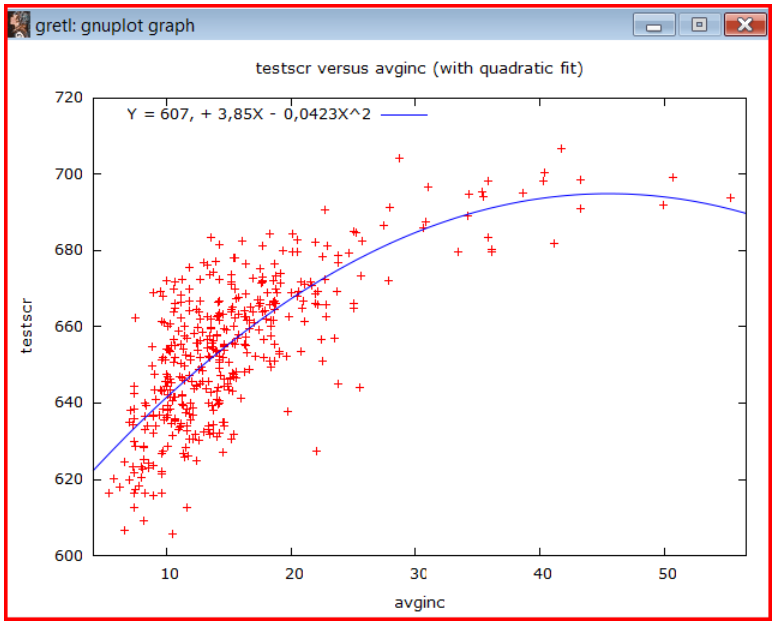
Resultatet är signifikant, vilket både t-värdena (som testar om $\beta_1 = 0$ och om $\beta_2 = 0$) och F-värdet (som testar om både β_1 och $\beta_2 = 0$) visar. Jämfört med den enkla regressionsmodelen har r^2 också stigit.

7.7 Kort om icke-linearitet

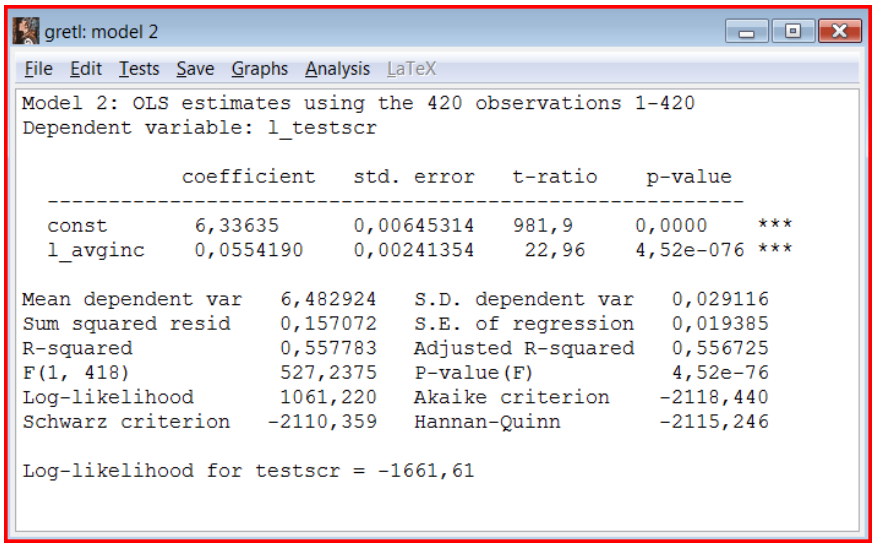
Som vi sett i diagrammen tidigare verkar det rimligt att beskriva relationen mellan testscr och str som linjär. Men om vi tittar på relationen testscr och avginc (genomsnittsinkomst) ser det annorlunda ut:



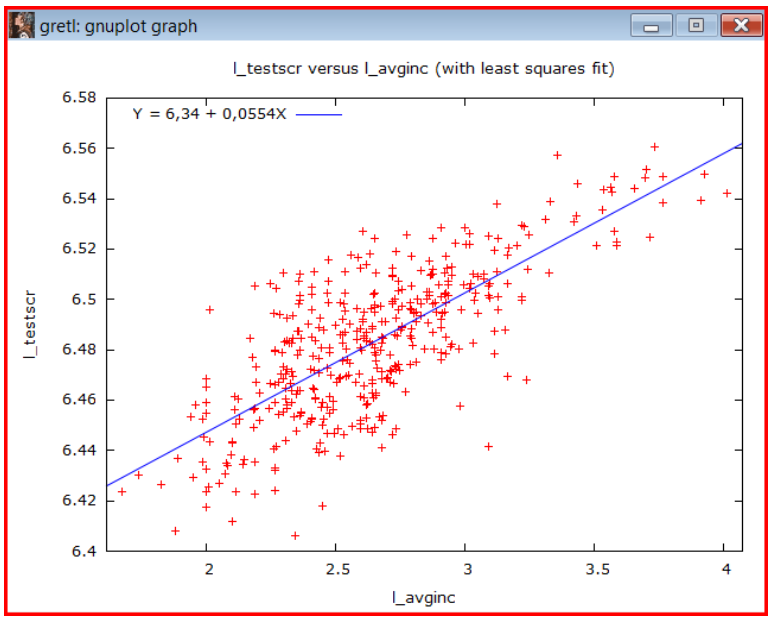
Om man likväl vill analysera relationen med hjälp av OLS kan man i princip gå tillväga på två olika sätt. Antingen skatta en kvadratisk (eller annan polynom) funktion och då får vi följande diagram



Eller så kan vi logaritmera variablerna. Markera testscr och avginc i huvudfönstret och välj sedan Add > Logs of selected variables. Då läggs de två nya variablerna l_avginc och l_testscr till längst ner i variabellistan i huvudfönstret. Välj sedan Model > Ordinary least squares Och för över l_testscr till beroende variabel och l_avginc till oberoende variabel. Tryck på OK och du får följande resultat:



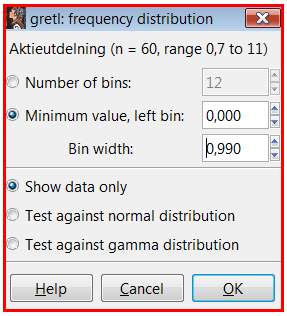
Välj sedan Save > Fitted values vilket lägger till yhat2 till variabellistan. Markera yhat2 och l_avginc och välj XY scatterplot med l_avginc som x-axelvariabel. Vi får



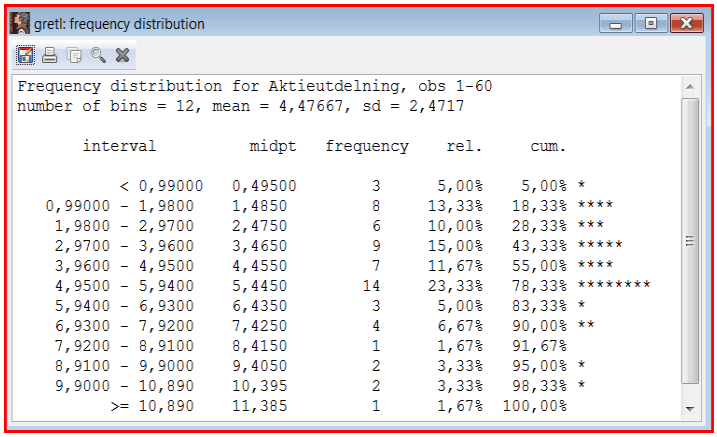
Här verkar det inte längre orimligt att bekriva relationen som linjär.

7.8 Frekvenstabeller

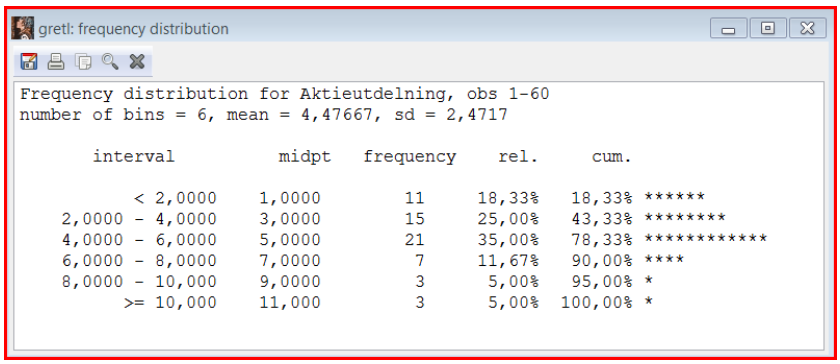
Ladda ner datafilen Tabell 4_1.gdt. För att få fram en frekvenstabell markerar du Aktieutdelning, klickar på Variable > Frequency distribution och fyller i:



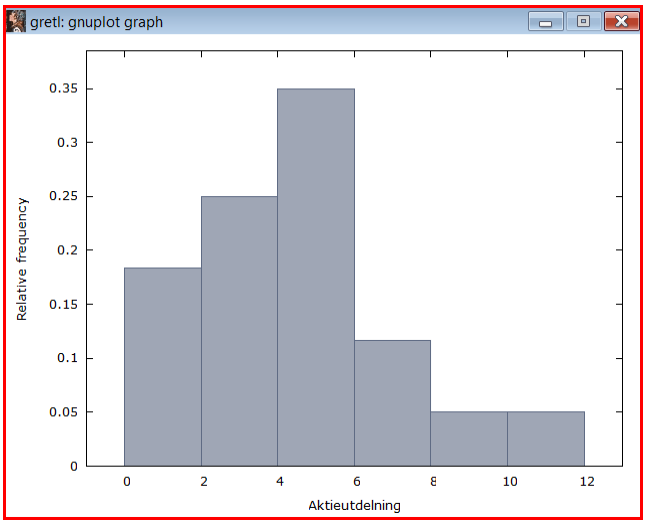
Tryck på OK och du får följande frekvenstabell:



Vill man ha större klassbredd, till exempel 2, får man i stället

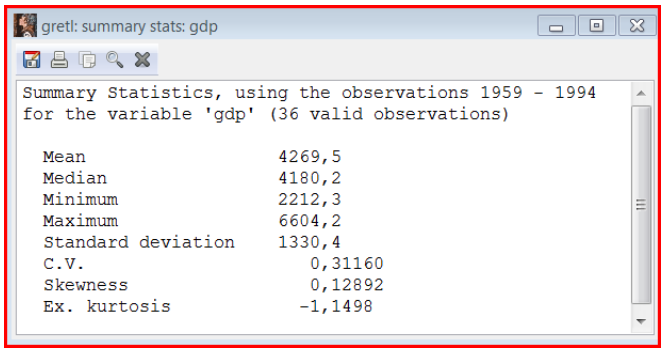


Och et frekvensdiagram får du genom att välja Variable > Frequency plot (och välj sedan som i det senaste exemplet ”minimumvalue, let bin” 0 och ”Bin width” 2) klicka på OK och du får diagrammet till höger. Notera att vi på y-axeln har relativfrekvensen (den absoluta finns i frekvenstabellen ovan).



7.9 Centralvärden

Det vanligaste värdet i en fördelning kallas **typvärdet**. Intressantare och vanligare i statistisk analys är median och medelvärde. **Medianen** är det värde för vilket det finns lika många observationer över som under detta värde. I **gretl** får vi enklast fram detta genom att markera en variabel, högerklicka och välja Descriptive statistics. För till exempel gdp i Ramanathan-mappens data 3_15.gdt får vi då:



Som vi ser är medianen 4180.2. Här ser vi också att medelvärdet är 4269.5. **Medelvärdet,** $\bar{X}$, beräknas med hjälp av formeln

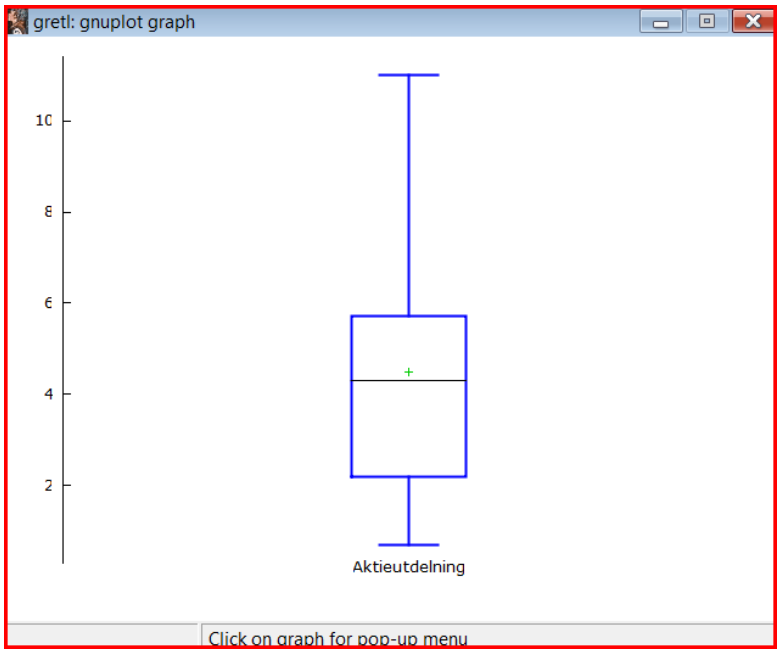
$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{N}$$

Om en fördelning är helt symmetrisk sammanfaller medianen och medelvärdet.

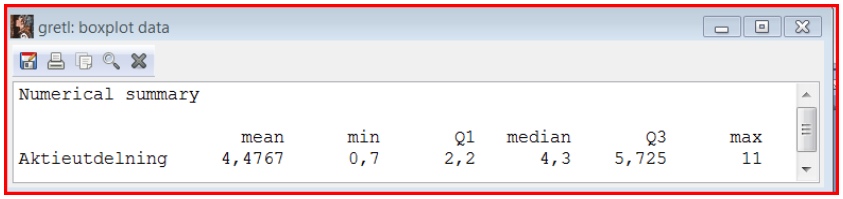
7.10 Spridningsmått

Om medianen delar en datamängd på mitten, så kan vi på liknande sätt definiera så kallade **kvartiler**. Första kvartilen är den punkt som har en fjärdedel av alla observationer under sig. På motsvarande sätt kan vi säga att tredje kvartilen har tre fjärdedelar av observationerna under sig.

I **gretl** finner vi kvartilerna lättast via boxplot. Markera till exempel variabeln Aktieutdelning i Tabell 4_1.gdt. Högerklicka sedan på variabeln och välj Boxplot i rullgardinen. Du får följande diagram:



Nedifrån och upp anges i diagrammet minimumvärde, första kvartil, medianvärde, medelvärde, tredje kvartil, maxvärde. Vill man ha dem klart utskrivna högerklickar man i diagrammet, väljer Numerical summary och får upp:



Spridningsmåttet **kvartilavvikelsen,** Q, är hälften av de mellersta 50 % av observationerna. Utifrån **gretl**s beteckningar kan det anges som $Q = \frac{Q_3 - Q_1}{2}$. I vårt exempel får vi då $Q = \frac{5.725 - 2.2}{2} = 1.76$. Om detta är "stort" eller "litet" vet vi inte förrän vi har något att jämföra med.

Ett annat spridningsmått är **variationsvidden**, som helt enkelt anger skillnaden mellan maximum- och minimumvärdena. I vårt exempel är den 6.52.

Det vanligaste spridningsmåttet är **standardavvikelsen**, s. Den definieras som

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{N}}$$

där X är de observerade värdena, $\bar{X}$ medelvärdet och N antalet observationer. I **gretl**s Descriptive statistics ges automatiskt standardavvikelsen (se Standard deviation i outputfönstret). I vårt exempel med variabeln Aktieutdelning är den 2.47.

Ett annat vanligt spridningsmått är **variansen**, som helt enkelt är standardavvikelsen i kvadrat, s^2.

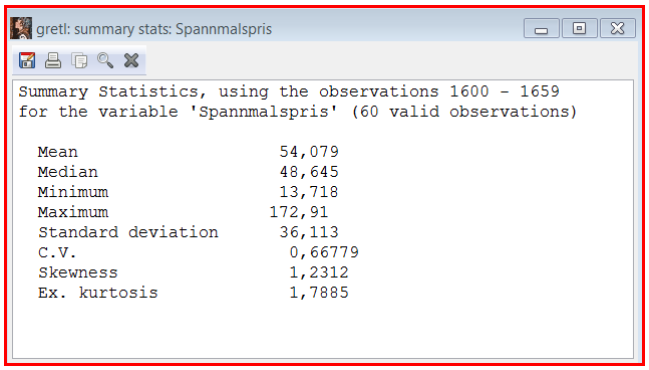
Ytterligare ett spridningsmått är **variationskoefficienten**, V, som vi definierar som

$$V = \frac{s}{\bar{X}}$$

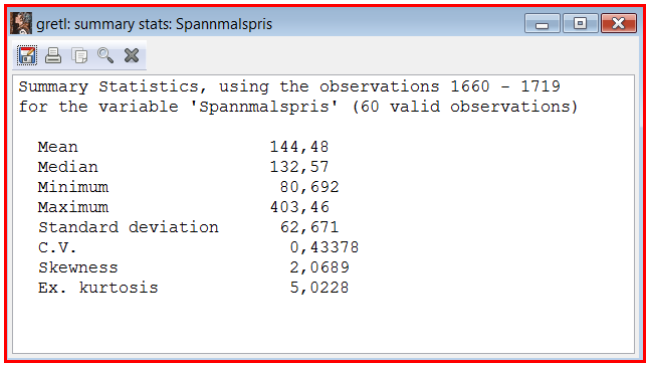
V är alltså standardavvikelsen i procent av medelvärdet. I **gretl** anges det som C.V. (som är förkortning för "coefficient of variation") i outputfönstret och är i vårt exempel 0.552 (multiplicerar vi talet med 100 får vi direkt procenttalet 55.2). Fördelen med V jämfört med standardavvikelsen är att den är oberoende av måttenheten. Överlag är det väl så att om det är absoluta skillnader vi är ute efter fungerar standardavvikelser bäst. Är det relativa skillnader vi är intresserade av är variationskoefficienten att föredra. Eftersom standardavvikelsen är känslig för extremvärden bör den dock som regel undvikas när vi har kraftigt asymmetriska fördelningar.

I Övning 6.4 i Eggeby och Söderberg ställs vi inför frågan om det skedde någon utjämning av spannmålspriserna under stormaktstiden i Sverige. Om vi öppnar datafilen Wage ex Sthlm 1600-1719.gdt och sedan i Sample sätter Restrict, based on criterion och skriver

Ar < 1660, har vi valt att ur datafilen välja att titta på perioden 1600-1669 (**gretl** går i sin automatiska datumsättning i GUI inte längre tillbaka än 1700, så vi har satt datum i konsolen genom kommandot setobs 1 1600 --time-series). Markera Spannmalspris och välj Descriptive statistics. Vi får då

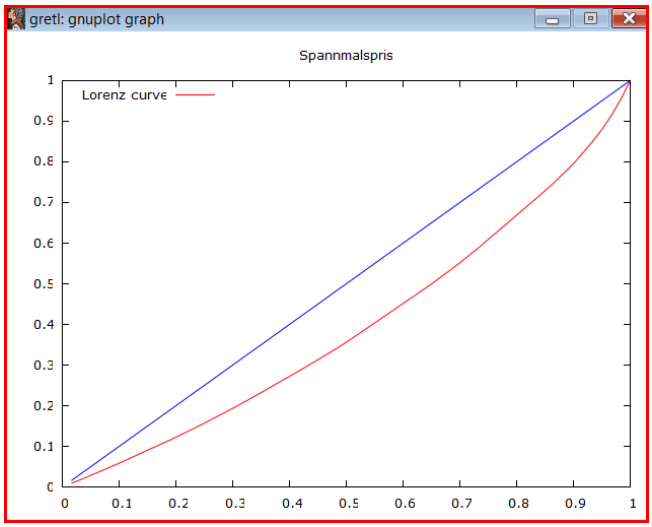


Gör vi om samma för åren 1660-1719 får vi

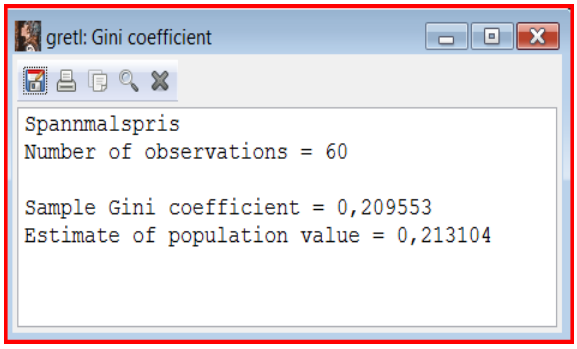


Vi ser att standardavvikelsen har ökat från 36.11 till 62.67 medan variationskoefficienten minskat från 0.67 till 0.43. Eftersom vi här är intresserade av relativa förändringar av prisvariationer över tiden är det senare måttet mest relevant och vi drar slutsatsen att prisvariationerna minskade under stormaktstiden.

Möjligen kan man här också tänka sig att använda **Ginikoefficienten** för att få ett mått på spridningen av priserna under de två tidsperioderna. I vanliga fall används Ginikoefficenten som ett ojämlikhetsmått på socialt och ekonomiskt område, men kan tillämpas även på detta sätt. Efter att ha valt tidsperiod, markera Spannmalspris, klicka på Variable och välj Gini coefficient i rullgardinen. För 1660-1719 får du följande diagram:



Och om du klickar bort diagrammet dyker följande upp:



Ginikoefficenten är 0.209. Gör vi motsvarande för perioden 1600-1659 får vi en Ginikoefficient på 0.348. Detta pekar i samma riktning som variations-koefficienten. Vi tolkar det som att marknadsintegration och andra orsaksfaktorer gjort prisvariationerna i Sverige mindre under stormaktstiden än under förutvarande period.