**Sommerskole i Videregående Statistik 2018**

**Opgave i regressionsanalyse**

**Formalia**

Opgaven besvares individuelt og uploades via Absalon (hvis det glipper så send en mail) fredag eftermiddag. Der skal et tekst dokument (word eller pdf) med selve besvarelsen, idet verbale fortolkninger og kommentarer er vigtige. Ved hvert delspørgsmål kan en stump af den anvendte SAS-kode med fordel copy pastes ind (gerne med farverne fra editoren) og det relevante output skal også copy pastes ind; men kun det output, der rent faktisk beskrives i teksten. Det samlede SAS-program copy-pastes ind sidst i denne tekstfil. I får hvert Jeres eget datasæt, jf beskrivelsen i mappen med datasæt.

**Problemstillingen**

Berlingske Tidendes erhvervssektions bagside fra 15/3-2001 var



Det fremgår altså at regressionsmodeller giver anledning til ballade - så er I advaret!

Den seriøse baggrund for artiklen er, at journalisterne har fået fat på en rapport fra Verdensbanken om virkningen af karteller på transportomkostningerne ved USA's import. Rapporten kan ses på Verdensbankens hjemmeside <http://www-wds.worldbank.org/> søg på WPS2522.

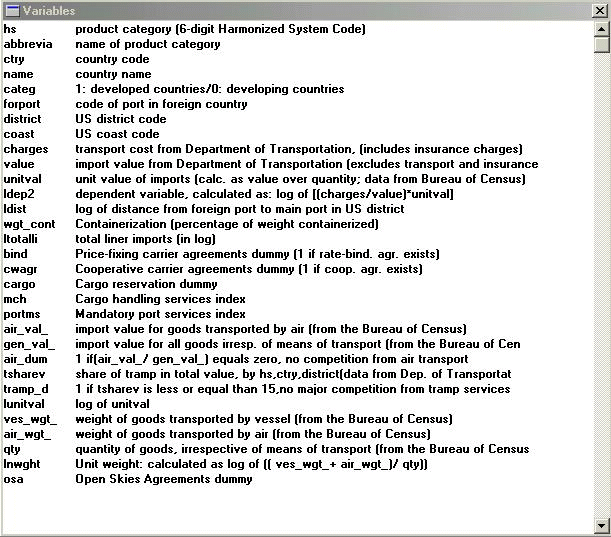
Balladeformlen er formel (5) side 14. I rapporten anvendes en model for alle skibstransporter til USA i året 1998 og der tages højde for en lang række forklarende variable (som fx. den distance, der sejles) inden der konkluderes, om dummyvariable for karteldannelse er signifikante. En af forfatterne (en tysker: Carsten Fink) har været venlig at sende hele datamaterialet til mig - i alt knapt 300.000 observationer.

Fra dette datamateriale har jeg udskilt en række varegrupper af typen A WOMEN'S OR GIRLS' TROUSERS ETC NOT KNIT, SYN FIBER. Dette navn fremgår af variablen **abbrevia** i datasættet, men denne variabel har en label **name of product category**, så I vil se denne mere sigende betegnelse i udskrifterne i stedet for variabelnavnet. I får et datasæt, der består af netop to af disse varer, dvs tal for alle skibstransporter for hver af de to varer.

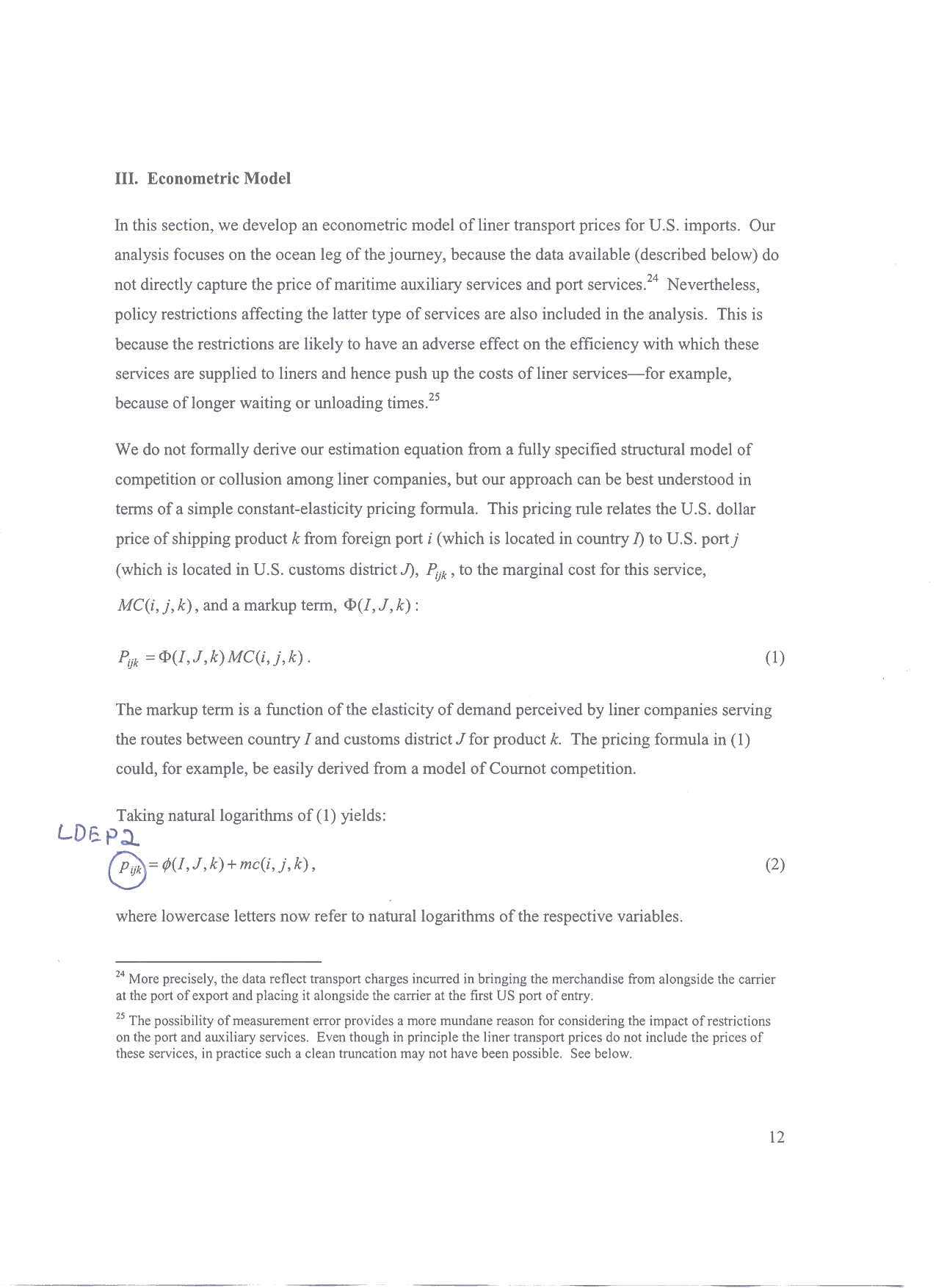
**Data**

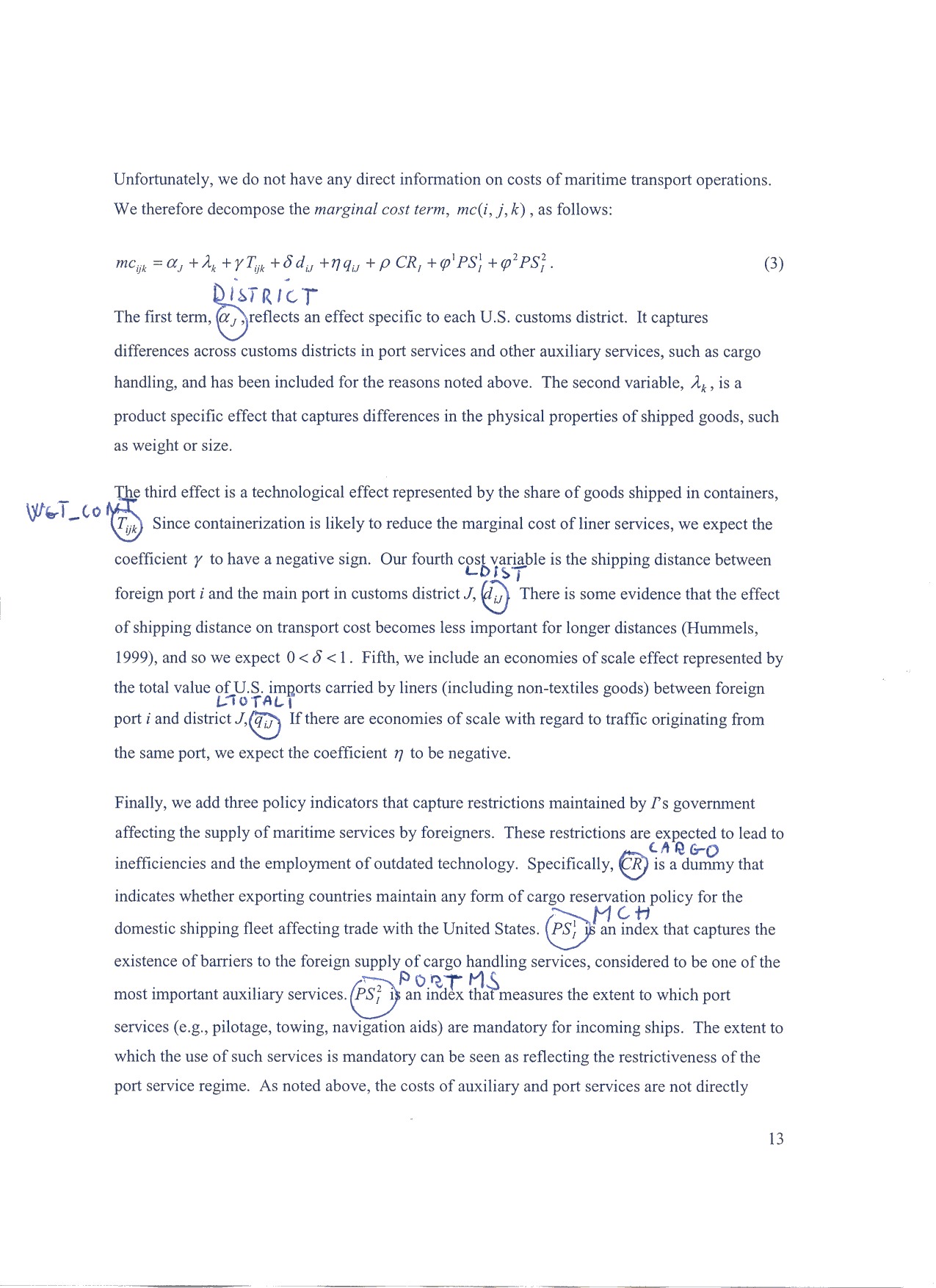
I Jeres datasæt er der en variabel **vare**, der lig med nul for den ene vare og lig med 1 for den anden vare.

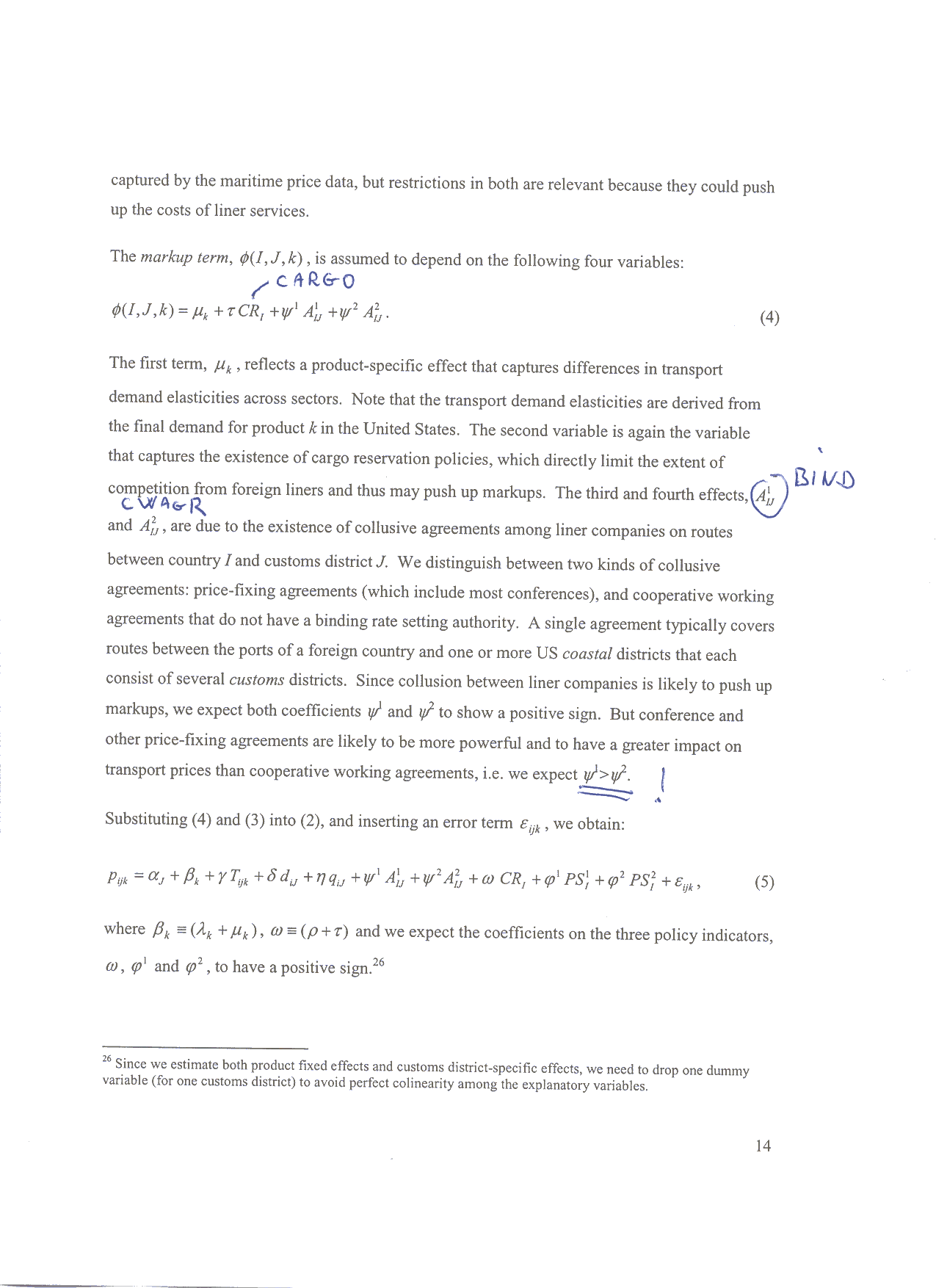
De øvrige variable fremgår af følgende tabel over variabelnavne (der skal bruges i programmeringen) og de mere signede labels, der anføres i udskrifter etc :

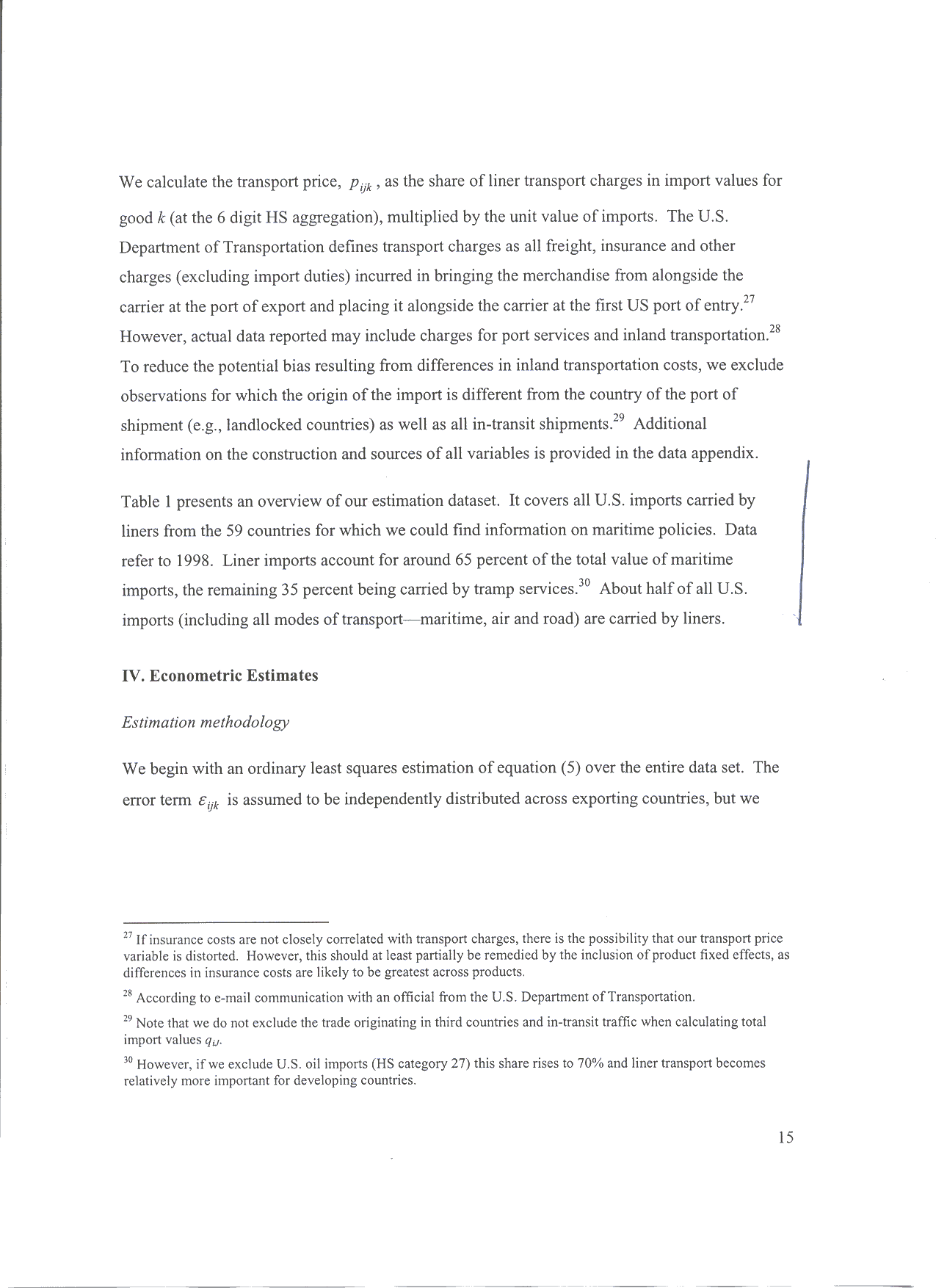


På de følgende sider er de relevante sider fra rapporten gengivet med håndskrevne henvisninger til disse variabelnavne.









Som udgangspunkt anvendes regressionsmodellen

**proc** **reg** data=???;

model ldep2=vare ldist wgt\_cont ltotalli bind cwagr cargo mch portms;

**run**;

**1)** Estimér modellens parametre. En modelkontrol er lang og uegnet til dokumentation, så af tidsmæssige grunde springes modelkontrollen over. Undersøg blot lineariteten i variablen **ldist** ved at inddrage dens kvadrat blandt de forklarende variable, for at se om der fx er faldende grænseomkostninger ved lange sejlture. Kvadratet udregnes ved

**data** k;

set ??;

**ldist\_kvadrat=ldist\*\*2;**

**run**;

**proc** **reg** data=k;

model ldep2=vare ldist **ldist\_kvadrat** wgt\_cont ltotalli bind cwagr cargo mch portms;

**run**;

**2)** Angiv en endelig model, hvori der kun er parametre, der er signifikante på 10% niveau. Gennemfør testet for at de øvrige forklarende variable simultant kan testes ud af modellen. Giv en kort, foreløbig fortolkning af den estimerede model relateret til problemstillingen - især omkring karteldannelsens betydning, dvs. de forklarende variable **bind** og **cwagr**. Tænk på, om der skal testes mod et ensidet alternativ for disse interessevariable.

**3)**

**a)** Gennemfør en analyse af de enkelte observationssæts indflydelse på estimationen.

**b)** Gennemfør et outliertest for det numerisk største eksternt standardiserede residual.

**c)** Undersøg ved hjælp af COVRATIO om eventuelle dårlige observationssæt skal udelades af den videre analyse.

**d)** Tegn de tilføjede variabel diagrammer (partial regression plots) og benyt også dem til en identifikation af grupper af outliere og grupper af indflydelsesrige sæt.

**4)** Identificer afvigende og eventuelt dårlige observationssæt ud fra resultatet af en LTS estimation på alle observationer. Prøv at fjerne "outliere" og "leverage punkter" hver for sig og samtidigt i en OLS-analyse med Proc Reg på de resterende "pæne" observationer. Sammenhold konklusionerne med konklusionerne fra partial regression plots (tilføjede variabel diagrammer) og analysen af enkeltobservationers indflydelse i opgave 3. Diskuter atter signifikansen af de forklarende variable i det nye lys. Ændres konklusionen omkring karteldannelsens betydning? **Da resultatet kan ændre på signifikansvurderingerne fra tidligere i opgavne, skal I begynde med den fulde model.** **Da LTS ikke er helt god til dummyvariable kan det være at LTS kørslen skal foregå uden dummy variable blandt de forklarende variable - det udelukker ikke, at dummyvariablene kan anvendes igen i en afsluttende OLS regression på de identificerede gode observationer.**

**5)** Nu er det jo ikke sikkert, at de forklarende variable betyder det samme for transportprisen for begge varer i Jeres datasæt. Prøv derfor at indføre vekselvirkninger mellem variablen **vare** og de øvrige forklarende variable, for at se om de øger forklaringsgraden væsentligt. Brug evt Proc GLM til et hurtigt tjek om vekselvirkningerne er signifikante. Giver fortegnene for de signifikante vekselvirkninger mening? Ændres konklusionen omkring karteldannelsens betydning?

**6)** I ovenstående antages, at effekten af prisaftaler kan ses ved en forhøjet pris på fragten. Omvendt kunne man tænke sig, at prisaftaler kunne opdages af konkurrencemyndighederne ud fra en forhøjet pris. Det svarer til modeller, hvor de variable **bind** og **cwagr** er responsvariable og prisen **ldep2** indgår blandt de forklarende variable. Da **bind** og **cwagr** er dummyvariable, kan det analyseres ved to adskilte logistiske regressioner. Gør det og konkludér, bl.a.ved brug af meningsfyldte effectplots. Vurder enkeltobservationers indflydelse på parameterestimaterne.

**7)** I spørgsmål 6) blev der bedt om to adskilte logistiske regressioner med hver sin responsvariabel. Prøv at kombinere dem ved at anvende variablen **aftale**, som defineret i det lille data-step nedenfor, som responsvariabel med flere end to kategorier i en logistisk regression med og uden parallelitetsforudsætningen. Prøv også en Poissonregression, selvom det næppe er en korrekt fordelingsantagelse i dette eksempel.

**data** a;

set ???;

aftale=**0**;

if bind = **1** and cwagr=**0** then aftale=**1**;

if bind = **0** and cwagr=**1** then aftale=**2**;

if bind = **1** and cwagr=**1** then aftale=**3**;

**run**;