EKSAMEN/EXAM STA-2000

Eksamen i/Examination in	**: STA-2000 Brukerkurs i statistikk 2**
Dato/Date	**:Tirsdag 7. desember 2005**
Tid/Time	**:Kl 09:00 til 13:00**
Sted/Site	**:Samvirkegården 4. etasje**
Tillatte hjelpemidler/Aids allowed	**:Alle skriftlige samt kalkulator/**
	All notes, books and pocket calculator

Oppgavesettet er på 12 sider inkludert førstesiden/
The exam paper consists of 12 pages including the front page

Kontaktperson under eksamen/ Contact person during the exam:
Amanuensis Elinor Ytterstad, tlf 77 64 4015

Eksamenssettet består av oppgave 1 a-g og oppgave 2 a-c /

The exam paper consists of Problem 1 a-g and Problem 2 a-c

Oppgave/Problem 1

Hvordan er drivstoff-forbruk (MPG) knyttet til de ulike kjøretøyvariablene? Datamaterialet (se tabell 1) er fra 1982 og inneholder observasjoner fra n=392 bilmodeller og følgende åtte variable: /

How is fuel consumption (MPG) connected to the various vehicle variables? The data set (see table 1) is from 1982 and contains observations from n=392 car models on the following eight variables:

MPG (Y, miles per gallon)
Cylinders (X_1, number of cylinders)
Engine Disp (X_2, engine displacement in cubic inches/ motorvolum)
Horsepower (X_3)
Weight (X_4, vehicle weight in pounds)
Accelerate (X_5, time to accelerate from 0 to 60mph in seconds)
Year (X_6, model year)
Origin (origin of car: American, European or Japanese)

Regresjonsmodellene 1 til 6 samt SPSS (statistikkprogram) resultater presenters etter delspørsmål g)./

The regression models 1 through 6 and SPSS (a statistical program) computer outputs are presented after question g).

a) På bakgrunn av korrelasjoner (tabell 2) og spredningsplott (figur 1), hvilke variable ser ut til å forklare responsen MPG? Begrunn svaret. /

   Based on the correlations (table 2) and scatter plots (figure 1), which variables seems to be explaining the response MPG. Explain why you choose these variables.

b) Test hypotesen: $H_0 : \beta_5 = 0$ i modell 3 under. Bruk 5% signifikansnivå. Beregn R^2 i modell 3. /

   In model 3 below, test the hypothesis: $H_0 : \beta_5 = 0$ on a 5% significance level. Calculate R^2 in model 3.

c) Betrakt modell 1 og test om kvadratleddene kan fjernes fra modellen. Formulèr null-hypotesen og utfør testen på 5% nivå. /

   Consider model 1 and test whether the squared terms should be excluded from the model. State the null-hypothesis and perform the test on a 5% significance level.

d) Hvilken modell er best av 1, 2 og 3? Forklar hvorfor. Hva kan sies om denne gruppen modeller (modell 1, 2 og 3) ved å betrakte figur 2? /

Which model of 1, 2 or 3 is the best, and why? What can be said about this group of models (model 1, 2 and 3) by looking at figure 2?

e) Betrakt modell 5. Er det noen uteliggerobservasjoner? I tilfelle ja, hvilken type uteligger er de, og vil de gi problemer? /

Consider model 5. Are there any outliers? If yes, what kind of outliers are they, and will they be troublesome?

f) Velg beste modell blant modell 4, 5 og 6.
Hvilken responsvariabel ser ut til å være den beste, MPG or ln(MPG)? Begrunn svaret./

Choose the best model among model 4, 5 and 6.
Which response variable seems to be better, MPG or ln(MPG)? Explain why.

g) Test i modell 5 om det er forskjell mellom japanske og europeiske biler mht til drivstoff-forbruk. Bruk 5% signifikansnivå. /

Test in model 5 whether there is any difference in fuel consumption between Japanese and European cars. Use 5% significance level.

The following models are considered:

1. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 Z_{am} + \beta_8 Z_{eur} + \beta_9 X_2^2 + \beta_{10} X_3^2 + \beta_{11} X_4^2 + E$

2. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 Z_{am} + \beta_8 Z_{eur} + E$

3. $Y = \beta_0 + \beta_4 X_4 + \beta_6 X_6 + \beta_{11} X_4^2 + \beta_7 Z_{am} + \beta_3 X_3 + \beta_{10} X_3^2 + \beta_5 X_5 + E$

A new response variable is made: $Y^* = \ln(Y)$, and the following models with this response are also considered:

4. $Y^* = \beta_0 + \beta_4 X_4 + \beta_6 X_6 + \beta_7 Z_{am} + \beta_3 X_3 + E$

5. $Y^* = \beta_0 + \beta_4 X_4 + \beta_6 X_6 + \beta_7 Z_{am} + \beta_3 X_3 + \beta_2 X_2 + \beta_1 X_1 + \beta_5 X_5 + \beta_8 Z_{eur} + E$

6. $Y^* = \beta_0 + \beta_6 X_6 + \beta_4 X_4 + \beta_7 Z_{am} + \beta_1 X_1 + \beta_3 X_3 + \beta_2 X_2 + E$

with $Z_{am} = \begin{cases} 1 & \text{if american car model} \\ 0 & \text{else} \end{cases}$, $Z_{eur} = \begin{cases} 1 & \text{if european car model} \\ 0 & \text{else} \end{cases}$

Table 1: Car data (an extract from totally 392 car models)

MPG	Cylinders	Engine Disp	Horsepower	Weight	Accelerate	Year	Origin
15,0	8	390	190	3850	8,5	70	American
21,0	6	199	90	2648	15,0	70	American
19,0	6	232	100	2634	13,0	71	American
18,0	6	258	110	2962	13,5	71	American
17,0	8	304	150	3672	11,5	72	American
15,0	8	304	150	3892	12,5	72	American
13,0	8	350	155	4502	13,5	72	American
24,0	4	120	97	2489	15,0	74	Japanese
26,0	4	108	93	2391	15,5	74	Japanese
32,0	4	71	65	1836	21,0	74	Japanese
31,0	4	76	52	1649	16,5	74	Japanese
20,0	6	232	100	2914	16,0	75	American
15,0	6	258	110	3730	19,0	75	American
23,0	4	115	95	2694	15,0	75	European
23,0	4	120	88	2957	17,0	75	European
30,0	4	97	67	1985	16,4	77	Japanese
26,0	4	97	75	2265	18,2	77	Japanese
19,4	6	232	90	3210	17,2	78	American
18,1	6	258	120	3410	15,1	78	American
20,6	6	231	105	3380	15,8	78	American
38,1	4	89	60	1968	18,8	80	Japanese
29,8	4	134	90	2711	15,5	80	Japanese
22,4	6	231	110	3415	15,8	81	American
32,0	4	144	96	2665	13,9	82	Japanese
34,0	4	108	70	2245	16,9	82	Japanese

Table 2: Correlations

	MPG	Cylinders	Engine Disp	Horsepower	Weight	Accelerate	Year
MPG	1	-,778	-,805	-,778	-,832	,423	,581
Cylinders	-,778	1	,951	,843	,898	-,505	-,346
Engine Disp	-,805	,951	1	,897	,933	-,544	-,370
Horsepower	-,778	,843	,897	1	,865	-,689	-,416
Weight	-,832	,898	,933	,865	1	-,417	-,309
Accelerate	,423	-,505	-,544	-,689	-,417	1	,290
Year	,581	-,346	-,370	-,416	-,309	,290	1

Figure 1

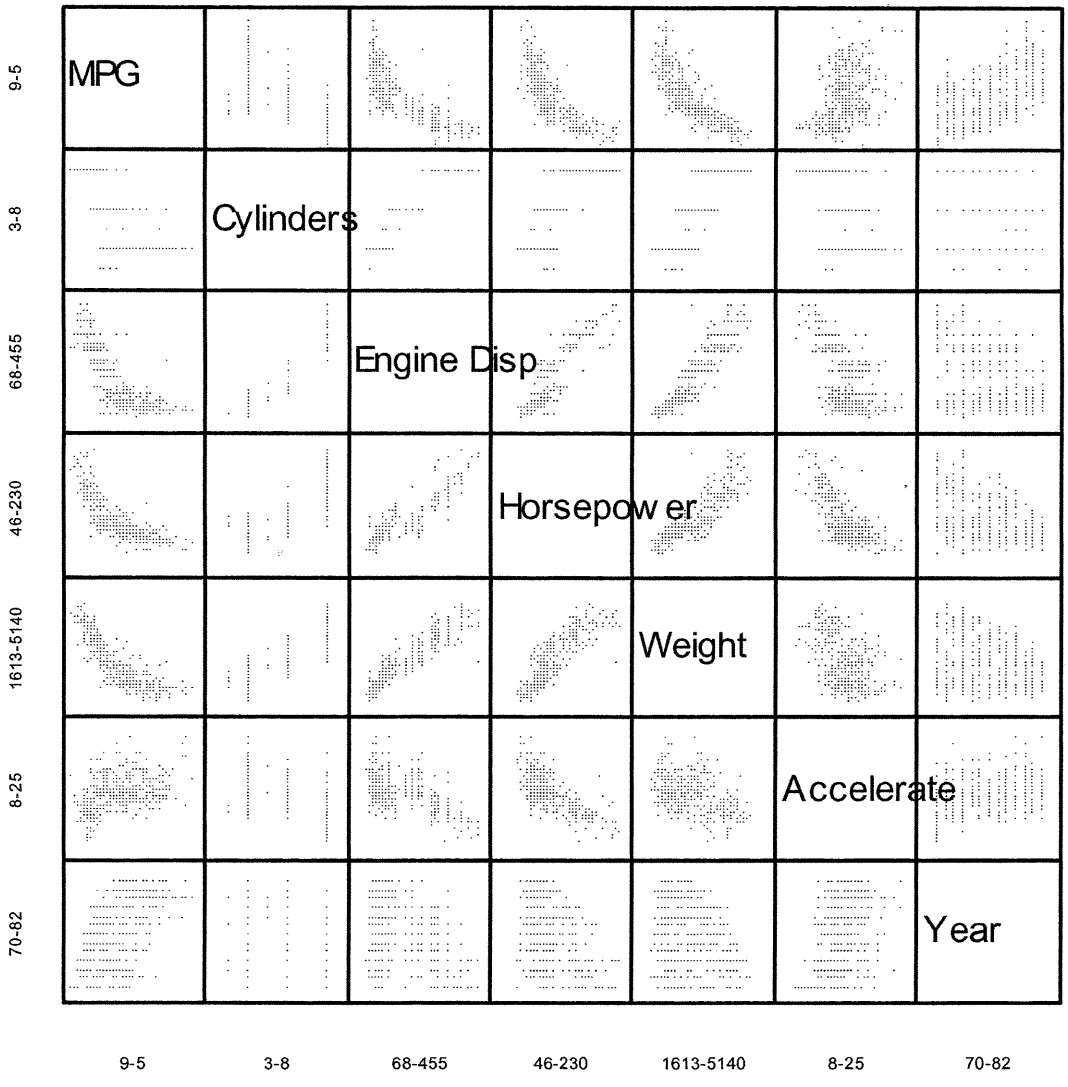


Table 3: A summary of results from SPSS

Model	SSY	SSE	df(SSE)	R^2	Mallow's C_p
1(full model)	23818.993	3196.568	380	0.866	12
2		4187.392	383	0.824	123.8
3		3222.794	384		7.1
4	45.210	5.487	387	0.879	10.2
5(full model)		5.358	383	0.881	9
6		5.361	385	0.881	5.2

Table 4 (model 3) Coefficients

		Unstandardized Coefficients			t	Sig.
Model		B	Std. Error			
3	(Constant)	7,321	4,460		1,641	,102
	Weight	-1,352E-02	,002		-6,982	,000
	Year	,784	,044		17,633	,000
	Weight*Weight	1,456E-06	,000		5,331	,000
	Zam	-1,322	,394		-3,359	,001
	Horsepower	-,199	,037		-5,332	,000
	Horse*Horse	5,944E-04	,000		5,147	,000
	Accelerate	-,188	,095		-1,979	,048

a Dependent Variable: MPG

Table 5 (model 4) Coefficients

		Unstandardized Coefficients			t	Sig.
Model		B	Std. Error			
4	(Constant)	1,695	,146		11,630	,000
	Weight	-2,500E-04	,000		-16,141	,000
	Year	3,013E-02	,002		16,637	,000
	Zam	-6,651E-02	,016		-4,263	,000
	Horsepower	-9,544E-04	,000		-2,908	,004

a Dependent Variable: LNY

Table 6 (model 5) Coefficients

		Unstandardized Coefficients			t	Sig.
Model		B	Std. Error			
5	(Constant)	1,793	,167		10,708	,000
	Weight	-2,639E-04	,000		-11,260	,000
	Year	3,055E-02	,002		16,491	,000
	Zam	-8,145E-02	,020		-4,119	,000
	Horsepower	-1,520E-03	,000		-3,100	,002
	Engine Disp	7,874E-04	,000		2,876	,004
	Cylinders	-2,781E-02	,011		-2,420	,016
	Accelerate	-1,403E-03	,004		-,399	,690
	Zeur	3,856E-03	,020		,190	,849

a Dependent Variable: LNY

Table 7 (model 6) Coefficients

		Unstandardized Coefficients			t	Sig.
Model		B	Std. Error			
6	(Constant)	1,770	,147		12,042	,000
	Year	3,054E-02	,002		16,869	,000
	Weight	-2,675E-04	,000		-13,270	,000
	Zam	-8,321E-02	,017		-4,832	,000
	Cylinders	-2,769E-02	,011		-2,422	,016
	Horsepower	-1,407E-03	,000		-3,678	,000
	Engine Disp	7,961E-04	,000		2,931	,004

a Dependent Variable: LNY

Figure 2, model 1

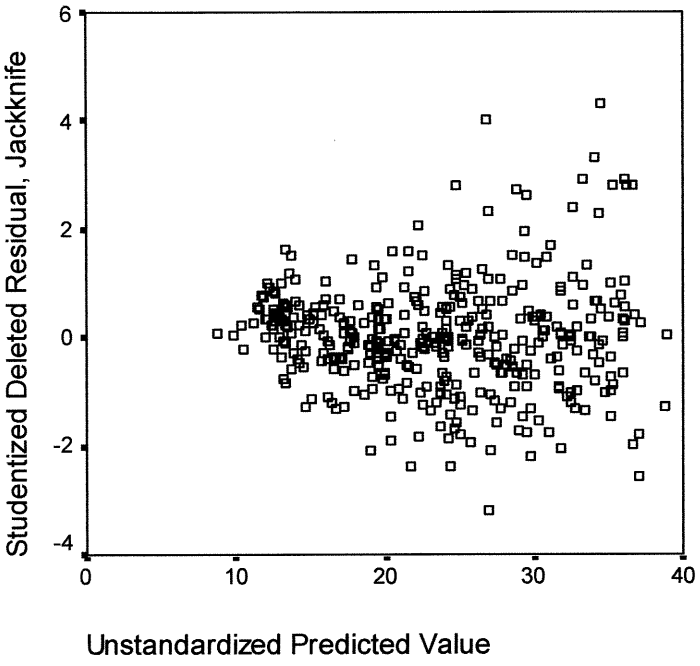


Figure 3, model 5

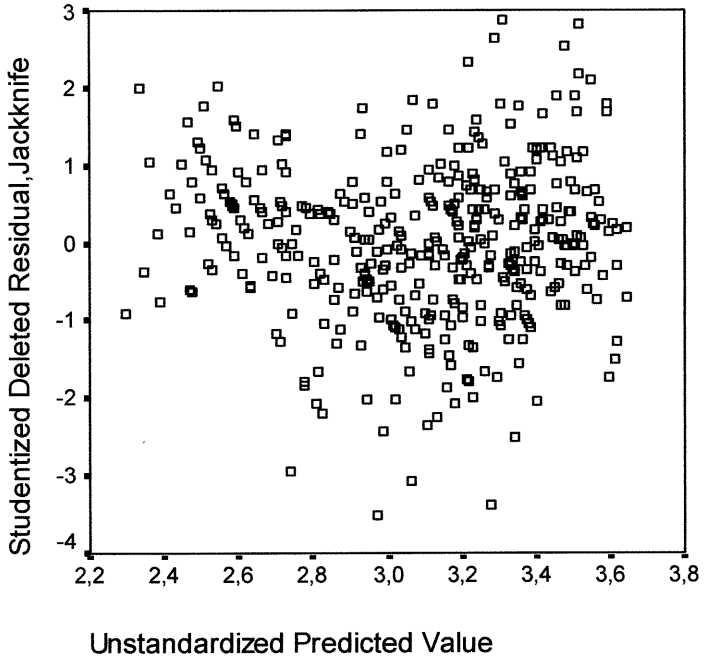


Figure 4, model 5

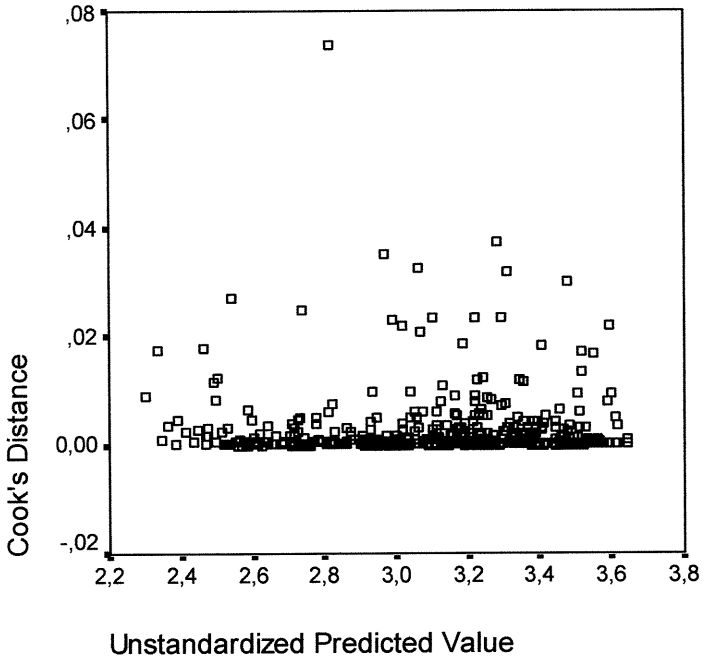
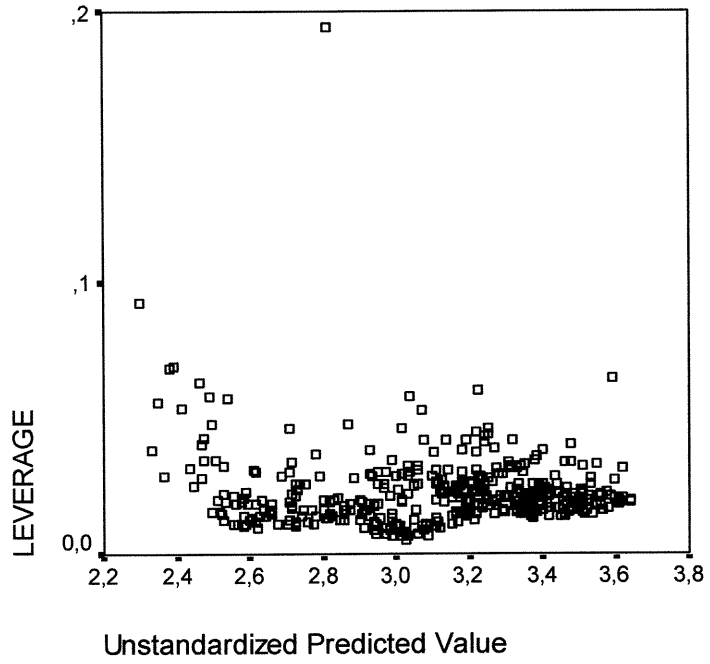


Figure 5, model 5



The leverage in figure 5 is calculated as defined in the textbook by Kleinbaum, Kupper, Muller & Nizam.

Oppgave/Problem 2

Er det dyrere å bo på hotell i New York City enn i andre storbyer i USA? For å svare på dette spørsmålet skal vi sammenligne rompriser i fire storbyer: Los Angeles (LA), New York City (NY), San Francisco (SF) and Washington D.C. (DC). To to-stjerners, to tre-stjerners og to fire-stjerners hoteller i hver by er valgt ut tilfeldig fra en reiseguidebok (tabell 1). Resultater er presentert i tabell 2 og figur 1. /

Are hotels in New York City more expensive then hotels in other major cities in USA? To answer this question let's look at room prices in four major cities: Los Angeles (LA), New York City (NY), San Francisco (SF) and Washington D.C. (DC). For each city a random sample of two two-stars, two three-stars and two four-stars hotels was taken from a travel guide book (table 1). Results are presented in table 2 and figure 1.

a) Er noen av faktorene *City* eller *Star* random faktorer i en variansanalyse? Begrunn svaret. /

   Are any of the factors *City* or *Star* random factors in an analysis of variance? Explain why/why not.

b) Er det signifikant interaksjon (samspill) mellom *City* og *Star*?
   Er det en signifikant prisforskjell mellom byer? Bruk 5% nivå. /

   Is there a significant interaction between City and Star?
   Is there a significant price difference between cities? Use a 5% level.

Ser av figur 1 at det kan være en lineær sammenheng mellom rompris og antall stjerner. Vi skal derfor også gjøre en regresjonsanalyse av dette datamaterialet (tabell 3 og 4).
La X=antall stjerner, Z_{dc}=1 hvis DC og 0 ellers, Z_{la}=1 hvis LA og 0 ellers, Z_{sf}=1 hvis SF og 0 ellers. /

Since hotel price seems to be linear connected to the number of stars (figure 1), we will also do a regression analysis of this data set (table 3 and 4).
Let X=number of stars, Z_{dc}=1 if DC and 0 else, Z_{la}=1 if LA and 0 else, Z_{sf}=1 if SF and 0 else.

c) Bruk regresjonsmodellen nedenfor og test om noen av de andre byene har signifikant forskjellig rompriser fra New York City når antall stjerner er tatt hensyn til i modellen. Bruk 5% signifikansnivå. /

   Use the linear regression model: $Y = \beta_0 + \beta_1 X + \beta_2 Z_{dc} + \beta_3 Z_{la} + \beta_4 Z_{sf} + E$ and test whether any of the other cities differ significantly from New York City in room price when number of stars are accounted for as in this linear model. Use a 5% significance level.

Table 1:

City	Hotel	Stars	Price	Z_{dc}	Z_{la}	Z_{sf}
DC	Lombardy	2	115	1	0	0
DC	Sheraton	2	185	1	0	0
DC	Hilton	3	166	1	0	0
DC	Grand Hyatt	3	189	1	0	0
DC	Park Hyatt	4	255	1	0	0
DC	Watergate	4	215	1	0	0
LA	Holiday Inn	2	79	0	1	0
LA	Le Dufy	2	145	0	1	0
LA	New Otani	3	119	0	1	0
LA	Hilton	3	150	0	1	0
LA	Biltmore	4	140	0	1	0
LA	Sofitel	4	160	0	1	0
NY	Helmsley Mid	2	135	0	0	0
NY	Holiday Inn	2	185	0	0	0
NY	The Mark	3	250	0	0	0
NY	Grand Hyatt	3	210	0	0	0
NY	The Peninsul	4	250	0	0	0
NY	The Regency	4	215	0	0	0
SF	Holiday Inn	2	99	0	0	1
SF	Diva	2	109	0	0	1
SF	Parc 55	3	170	0	0	1
SF	Inn at Opera	3	110	0	0	1
SF	Mandarin	4	265	0	0	1
SF	Grand Hyatt	4	169	0	0	1

Table 2: ANOVA-table from two-way analysis of variance

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	49552,458	11	4504,769		
Intercept	695301,042	1	695301,042		
CITY	20462,792	3	6820,931		
STARS	23794,083	2	11897,042		
CITY * STARS	5295,583	6	882,597	,684	,667
Error	15493,500	12	1291,125		
Total	760347,000	24			
Corrected Total	65045,958	23			

a R Squared = ,762

Figure 1: Hotel Price

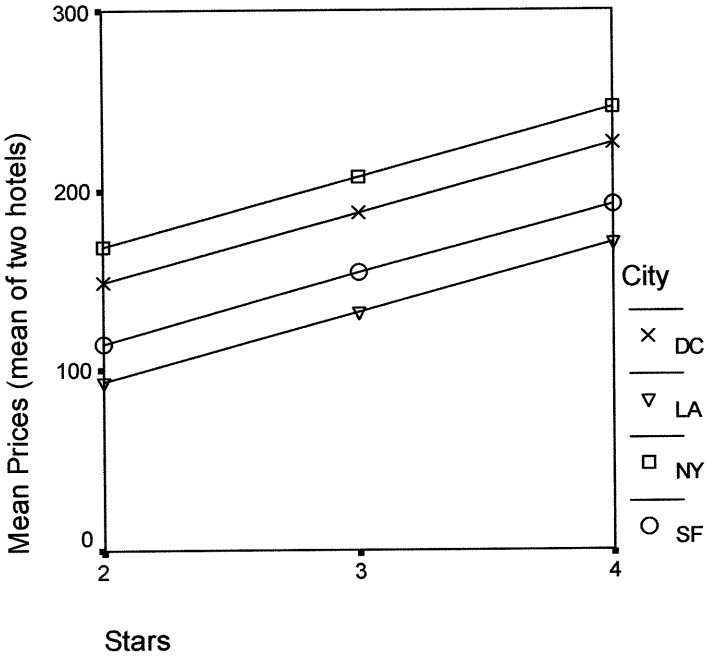


Table 3: ANOVA from a regression analysis

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	44255,854	4	11063,964	10,111	,000
	Residual	20790,104	19	1094,216		
	Total	65045,958	23			

a Predictors: (Constant), Zsf, Stars, Zla, Zdc
b Dependent Variable: Price

Table 4: Coefficients

		Unstandardized Coefficients			t	Sig.
Model		B	Std. Error			
1	(Constant)	91,813	28,247		3,250	,004
	Stars	38,563	8,270		4,663	,000
	Zdc	-20,000	19,098		-1,047	,308
	Zla	-75,333	19,098		-3,945	,001
	Zsf	-53,833	19,098		-2,819	,011

a Dependent Variable: Price