

Lineære modeller

Regresjonsmodeller med

én responsvariabel: Y

én eller flere "uavhengige" variabler: $x_1, \dots, x_k$

Den betingede fordeling av Y gitt $x_1, \dots, x_k$ antas at være **normal** med en forventning som avhenger av x-ene og en varians som er uavhengig av x-ene:

$$E(Y|x_1, x_2, \dots, x_k) = \beta_0 + \sum_{j=1}^k \beta_j x_j$$

- De uavhengige variablene kan være kategorivariabler eller kvantitative.
- Kategoriske variabler omkodes som dummyvariabler.
- Interaksjoner kan tilføyes i form av produktvariabler.

1

Forskjellige spesialtilfeller

Uavhengige variabler	Analysen omtales som
Én binær variabel	Toutvalgs t-test
Én kategorivariabel med flere enn to kategorier	Enveis variansanalyse
To eller flere kategorivariabler	To- eller flerveis variansanalyser
En eller flere kvantitative variabler	Lineær regresjon
Både kategorivariabler og kvantitative variabler	(Generelle) Lineære modeller (*)

*) Dekker også en rekke komplekse modeller, multivariabel respons og avhengige observasjoner

2

Lineære modeller i praksis

Eksempel:

Utføre en analyse av effekten av røykevaner og distraksjon på hyppigheten av feil gjort i utførelsen av visse oppgaver. Designet forsøk, **balansert** design.

Responsvariabel: Antall feil (errors)

Uavhengige: Oppgave (task)
Røykevaner (smkgrp)
Grad av distraksjon (distract)

3

Primært problem:

Hvilken betydning har de uavhengige variablene for feilhyppighet?

Er det **interaksjoner** (samspill) mellom effekten av de uavhengige variablene?

Sekundære problemer:

Er relasjonen mellom grad distraksjon og feilhyppighet **lineær**?

Kan den betingede fordeling av feilhyppighet og de andre variablene antas å være **normal**?

Er **variansene** **homogene**?

4

Analyseplan

Innledende databeskrivelse

- Univariat
- Bivariat

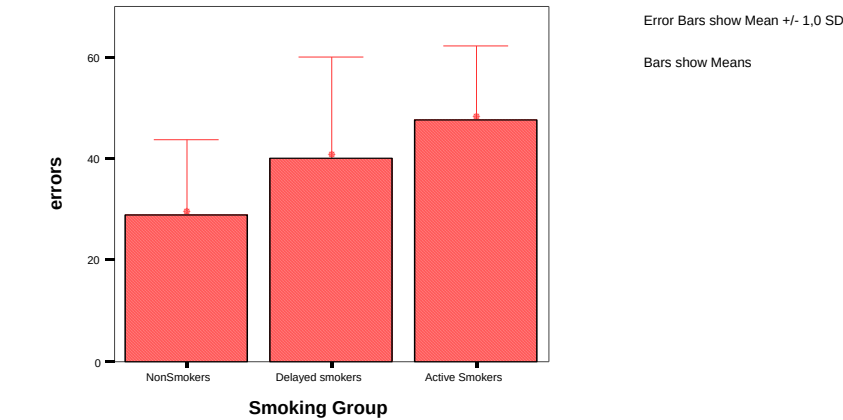
Den innledende analyse kan gi ideer til den egentlige analyse, men skal ikke i seg selv føre til endelige analyseresultater pga. risikoen for konfundering.

Den egentlige analyse:

- Trinnvis modellsøking
- Modellkontroll
- Estimering av parametrer i den endelige modell

task	smkgrp	errors	distract
1	1	1	9
1	1	1	8
...	...	...	...
1	1	1	10
1	2	12	101
1	2	7	75
...	...	...	...
1	2	16	118
1	3	8	64
1	3	8	135
...	...	...	...
1	3	10	123
2	1	27	126
2	1	34	154
...	...	...	...
2	1	49	143
2	2	48	113
2	2	29	100
...	...	...	...
2	2	49	96
2	3	34	108
2	3	65	191
...	...	...	...
2	3	47	132
3	1	15	110
3	1	2	96
...	...	...	...
3	1	13	116
3	2	7	93
3	2	0	102
...	...	...	...
3	2	11	102
3	3	3	130
3	3	2	83
...	...	...	...
3	3	3	114

Feilhyppighet i ”Task” = 2 fordelt på røykevaner



Descriptives

	errors		Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
	N	Mean			Lower Bound	Upper Bound		
NonSmokers	15	28,87	14,687	3,792	20,73	37,00	4	56
Delayed smokers	15	39,93	20,133	5,198	28,78	51,08	6	71
Active Smokers	15	47,53	14,652	3,783	39,42	55,65	21	75
Total	45	38,78	18,055	2,692	33,35	44,20	4	75

ANOVA

errors					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2643,378	2	1321,689	4,744	,014
Within Groups	11700,400	42	278,581		
Total	14343,778	44			

Kontraster

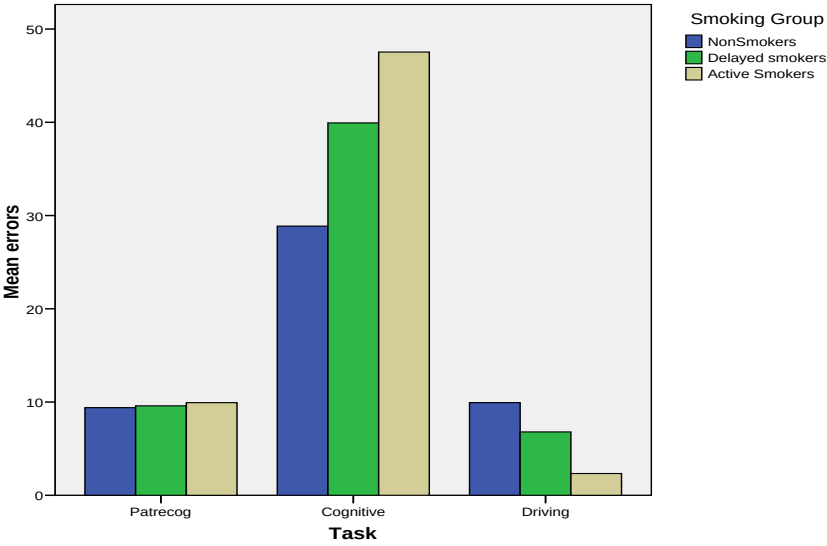
Contrast	Smoking Group		
	NonSmokers	Delayed smokers	Active Smokers
1	1	-,5	-,5
2	0	1	-1

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
errors	Assume equal variances	1	-14,87	5,278	-2,817	42	,007
		2	-7,60	6,095	-1,247	42	,219
	Does not assume equal variances	1	-14,87	4,971	-2,990	32,240	,005
		2	-7,60	6,429	-1,182	25,581	,248

task	smkgrp	errors	distract
1	1	1	107
1	1	1	133
...	...	...	...
1	1	1	10
1	2	12	101
1	2	7	75
...	...	...	...
1	2	16	118
1	3	8	64
1	3	8	135
...	...	...	...
1	3	10	123
2	1	27	126
2	1	34	154
...	...	...	...
2	1	49	143
2	2	48	113
2	2	29	100
...	...	...	...
2	2	49	96
2	3	34	108
2	3	65	191
...	...	...	...
2	3	47	132
3	1	15	110
3	1	2	96
...	...	...	...
3	1	13	116
3	2	7	93
3	2	0	102
...	...	...	...
3	2	11	102
3	3	3	130
3	3	2	83
...	...	...	...
3	3	3	114

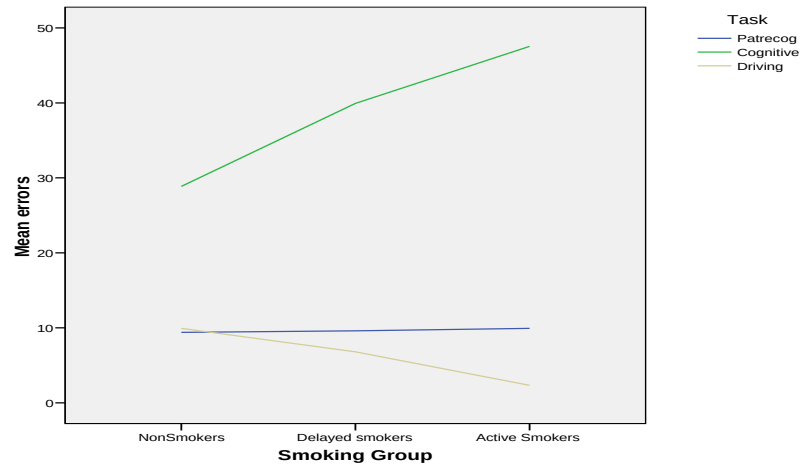
Feilhyppighet fordelt på oppgaver og røykevaner



Descriptive Statistics

Dependent Variable: errors				
Task	Smoking Group	Mean	Std. Deviation	N
Patrecog	NonSmokers	9,40	1,404	15
	Delayed smokers	9,60	4,405	15
	Active Smokers	9,93	6,519	15
	Total	9,64	4,513	45
Cognitive	NonSmokers	28,87	14,687	15
	Delayed smokers	39,93	20,133	15
	Active Smokers	47,53	14,652	15
	Total	38,78	18,055	45
Driving	NonSmokers	9,93	6,006	15
	Delayed smokers	6,80	5,441	15
	Active Smokers	2,33	2,289	15
	Total	6,36	5,701	45
Total	NonSmokers	16,07	12,828	45
	Delayed smokers	18,78	19,359	45
	Active Smokers	19,93	21,975	45
	Total	18,26	18,393	135

Samspill



13

Toveis ANOVA-modell

$$error_{ijk} = \mu + task_i + smkgrp_j + task \cdot smkgrp_{ij} + \varepsilon_k$$

Tests of Between-Subjects Effects

Dependent Variable: errors

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	31744,726(a)	8	3968,091	36,798	,000
Intercept	45009,074	1	45009,074	417,389	,000
smkgrp	354,548	2	177,274	1,644	,197
task	28661,526	2	14330,763	132,895	,000
smkgrp * task	2728,652	4	682,163	6,326	,000
Error	13587,200	126	107,835		
Total	90341,000	135			
Corrected Total	45331,926	134			

a. R Squared = ,700 (Adjusted R Squared = ,681)

14

Parameter Estimates

Hva betyr alt dette?

Dependent Variable: errors

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	2,333	2,681	,870	,386	-2,973	7,639
[smkgrp=1]	7,600	3,792	2,004	,047	,096	15,104
[smkgrp=2]	4,467	3,792	1,178	,241	-3,037	11,971
[smkgrp=3]	0(a)	.	.	.	.	.
[task=1]	7,600	3,792	2,004	,047	,096	15,104
[task=2]	45,200	3,792	11,920	,000	37,696	52,704
[task=3]	0(a)	.	.	.	.	.
[smkgrp=1] * [task=1]	-8,133	5,362	-1,517	,132	-18,745	2,479
[smkgrp=1] * [task=2]	-26,267	5,362	-4,898	,000	-36,879	-15,655
[smkgrp=1] * [task=3]	0(a)	.	.	.	.	.
[smkgrp=2] * [task=1]	-4,800	5,362	-,895	,372	-15,412	5,812
[smkgrp=2] * [task=2]	-12,067	5,362	-2,250	,026	-22,679	-1,455
[smkgrp=2] * [task=3]	0(a)	.	.	.	.	.
[smkgrp=3] * [task=1]	0(a)	.	.	.	.	.
[smkgrp=3] * [task=2]	0(a)	.	.	.	.	.
[smkgrp=3] * [task=3]	0(a)	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

15

Omkoding av kategorisk variabel "Task"

$$t_1 = \begin{cases} 1, \text{Task}=1 \text{ ("Patrecog")} \\ 0, \text{ellers} \end{cases}$$

$$t_2 = \begin{cases} 1, \text{Task}=2 \text{ ("Cognitive")} \\ 0, \text{ellers} \end{cases}$$

Konsekvens: Dersom Task =3 ("Driving"), blir t_1 og t_2 begge lik 0.

16

Omkoding av kategorisk variabel "Smoke"

$$s_1 = \begin{cases} 1, \text{Smoke}=1 \text{ ("Nonsmokers")} \\ 0, \text{ellers} \end{cases}$$

$$s_2 = \begin{cases} 1, \text{Smoke}=2 \text{ ("Delayed smokers")} \\ 0, \text{ellers} \end{cases}$$

Konsekvens: Dersom Smoke =3 ("Active smokers"), blir s_1 og s_2 begge lik 0.

17

Dummyvariabler for samspill

Samspill SMKGRP*TASK:

4 nye indikatorvariabler: $t1s1=t1*s1$, $t1s2=t1*s2$, $t2s1=t2*s1$, $t2s2=t2*s2$

Samspill opptrer når effekten av en faktor (f.eks. røykevaner) varierer med nivået til en annen faktor (her TASK). Rent konkret kunne det antas at røyking gav større (el. mindre) feilhyppighet på en kognitiv oppgave, og omvendt på bilkjøring.

18

Regresjonsmodell

$$y = \alpha + \beta_1 t1 + \beta_2 t2 + \beta_3 s1 + \beta_4 s2 + \beta_5 t1s1 + \beta_6 t1s2 + \beta_7 t2s1 + \beta_8 t2s2 + \varepsilon$$

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	31744,726	8	3968,091	36,798	,000(a)
	Residual	13587,200	126	107,835		
	Total	45331,926	134			

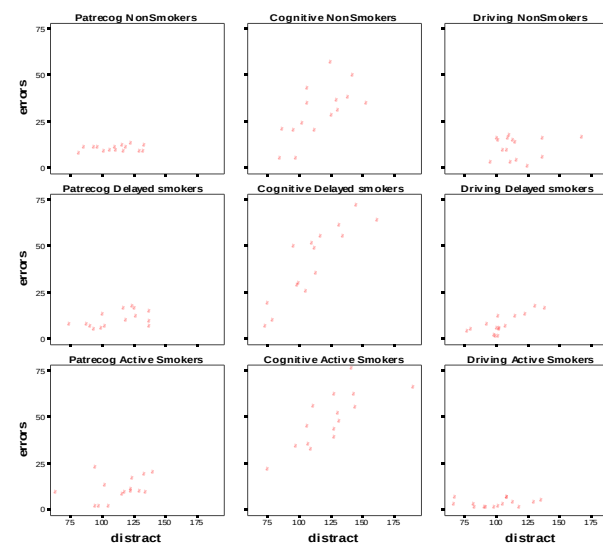
a Predictors: (Constant), t2s2, t2s1, t1s2, t1s1, t2, t1, s2, s1

b Dependent Variable: errors

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,333	2,681		,870	,386
	t1	7,600	3,792	,196	2,004	,047
	t2	45,200	3,792	1,163	11,920	,000
	s1	7,600	3,792	,196	2,004	,047
	s2	4,467	3,792	,115	1,178	,241
	t1s1	-8,133	5,362	-,139	-1,517	,132
	t1s2	-4,800	5,362	-,082	-,895	,372
	t2s1	-26,267	5,362	-,450	-4,898	,000
	t2s2	-12,067	5,362	-,207	-2,250	,026

19



20

Regresjonsmodell

$$y = \alpha + \beta_1 t_1 + \beta_2 t_2 + \beta_3 s_1 + \beta_4 s_2 + \beta_5 t_1 s_1 + \beta_6 t_1 s_2 + \beta_7 t_2 s_1 + \beta_8 t_2 s_2 + \delta \cdot \text{distract} + \varepsilon$$

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36389,602	9	4043,289	56,519	,000(a)
	Residual	8942,324	125	71,539		
	Total	45331,926	134			

a. Predictors: (Constant), distract, t1, s1, t2s2, t1s2, t2s1, s2, t1s1, t2

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-27,093	4,255		-6,367	,000
	t1	3,739	3,125	,096	1,196	,234
	t2	37,965	3,216	,977	11,804	,000
	s1	2,686	3,148	,069	,853	,395
	s2	3,102	3,093	,080	1,003	,318
	t1s1	-2,439	4,425	-,042	-,551	,582
	t1s2	-2,928	4,374	-,050	-,669	,504
	t2s1	-18,700	4,468	-,321	-4,186	,000
	t2s2	-6,450	4,423	-,111	-1,458	,147
	distract	,293	,036	,337	8,058	,000

21

Forlengs seleksjon av regressorer

Coefficients(a)

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8,000	1,189		6,730	,000
	t2	30,778	2,059	,792	14,949	,000
2	(Constant)	-26,236	4,302		-6,099	,000
	t2	28,469	1,708	,732	16,670	,000
	distract	,311	,038	,359	8,170	,000
3	(Constant)	-25,753	3,906		-6,593	,000
	t2	33,273	1,787	,856	18,619	,000
	distract	,307	,035	,354	8,871	,000
	t2s1	-14,315	2,649	-,245	-5,404	,000

a. Dependent Variable: errors

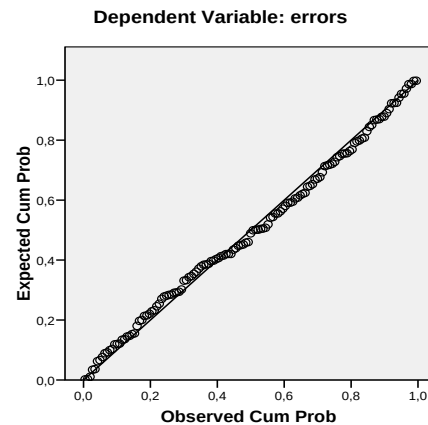
Forslag til endelig modell:

$$y = -23.753 + 33.273t_2 - 14.315t_2s_1 + 0.307 \cdot \text{distract} + \varepsilon$$

22

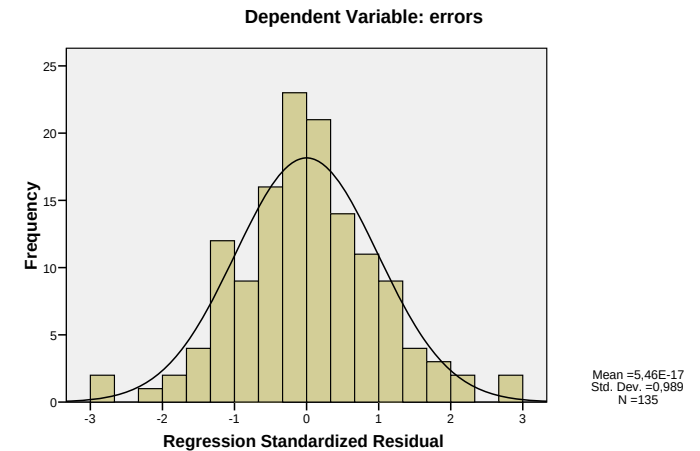
Sjekk av endelig modell

Normal P-P Plot of Regression Standardized Residual



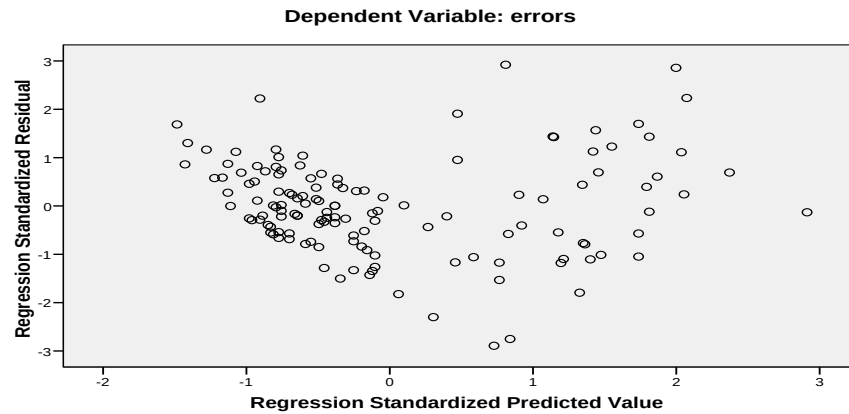
23

Histogram



24

Scatterplot



Homogen varians? Transformere?

Utliggere? Ekskludere?

Endelig konklusjon

Uansett røykevane er det ingen signifikant forskjell på feilhyppigheten i henholdsvis mønstergjenkjenning og bilkjøring.

Sammenliknet med bilkjøring er feilhyppigheten i kognitiv test større med en faktor på 33.

Sammenliknet med aktiv røyker reduseres feilhyppigheten i en kognitiv test med en faktor på -14 hos en ikke-røyker.

Grad av distraksjon øker feilhyppigheten med en faktor på 0.3 gitt oppgave og røykevane.