

Poissonregression och "count data"

- analyser då utfallet är antal händelser inom ett tidsintervall

Anna Lindam, statistiker

FoU – enheten Region Jämtland Härjedalen

Till Forum Norrs lunchföreläsning i metod 4 maj 2021

"count data"

- Antal gånger ett event sker inom ett visst tidsintervall
- Positiva heltal

Vad är en poissonregression?

- Regressionsmodell - generaliserad linjär modell
- Predicera utfallet med en eller flera parametrar

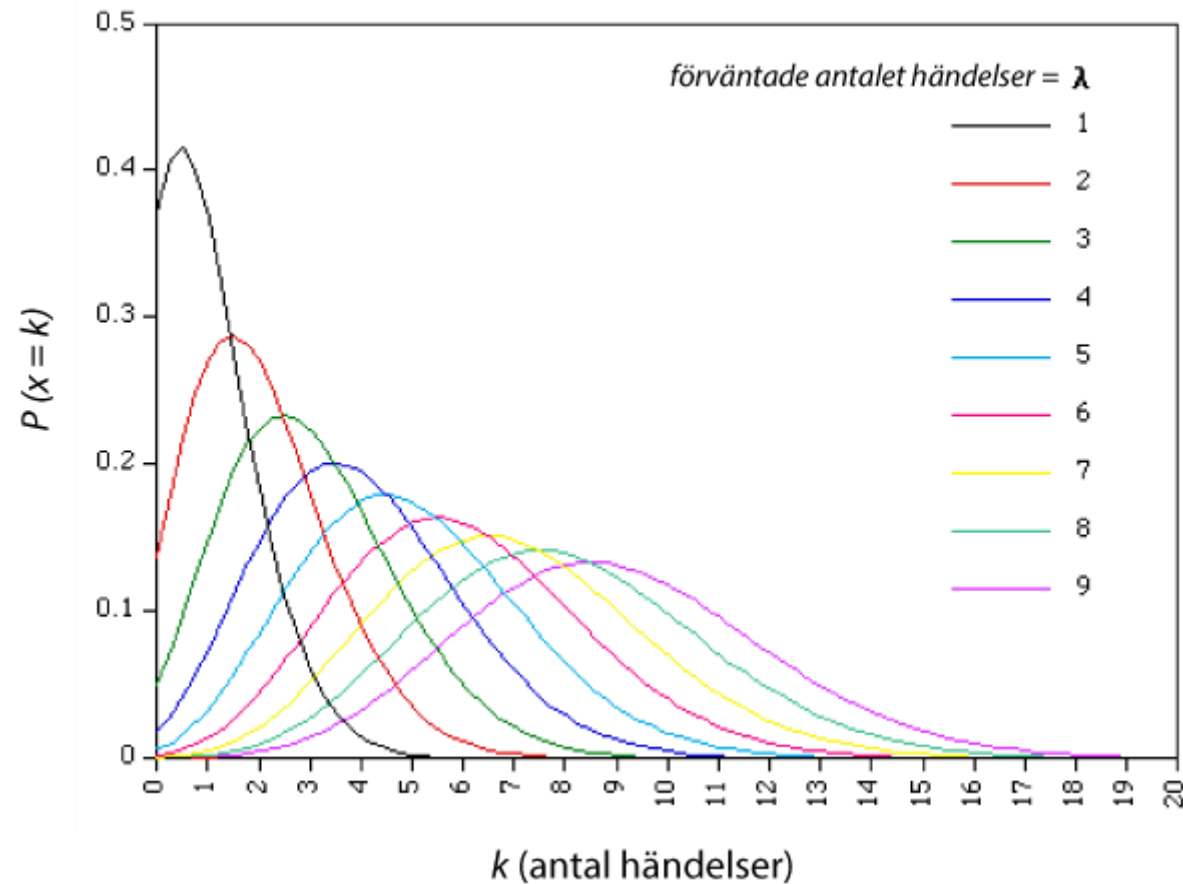
När används poissonregression?

- När utfallet är en antal events som sker under ett visst tidsintervall
- När händelserna sker oberoende av varandra
- Flexibel regressionstyp, kan även användas för
 - Överlevnadsanalys
 - Logistisk regression

Exempel frågeställningar

- Vad är sambandet mellan halsbränna och sömntablettsanvändning i dagar?
- Vad predicerar antal besök per år på olika hälsocentraler?
- Vilka faktorer påverkar antal sjukhusinläggningar sedan första diagnos?
- Finns det något samband mellan storlek på bostadsort och antal gymbesök?

Poissonfördelningen



Siméon Denis Poisson
Matematiker
1781 - 1840

Ett exempel

Som en del i en hälsoundersökning fick 156 personer ett gratis gymkort under 6 månader. Studiedeltagarna följdes sedan upp och man ville veta om det fanns ett samband mellan storleken på bostadsorten och hur många gånger personerna besökte gymmet under 6-månadersperioden.

Frågeställning: Finns det ett samband mellan storleken på bostadsort och hur ofta man tränar på gym?

Hypotes: Personer som lever i mer tätbefolkade områden tränar oftare.

Antaganden

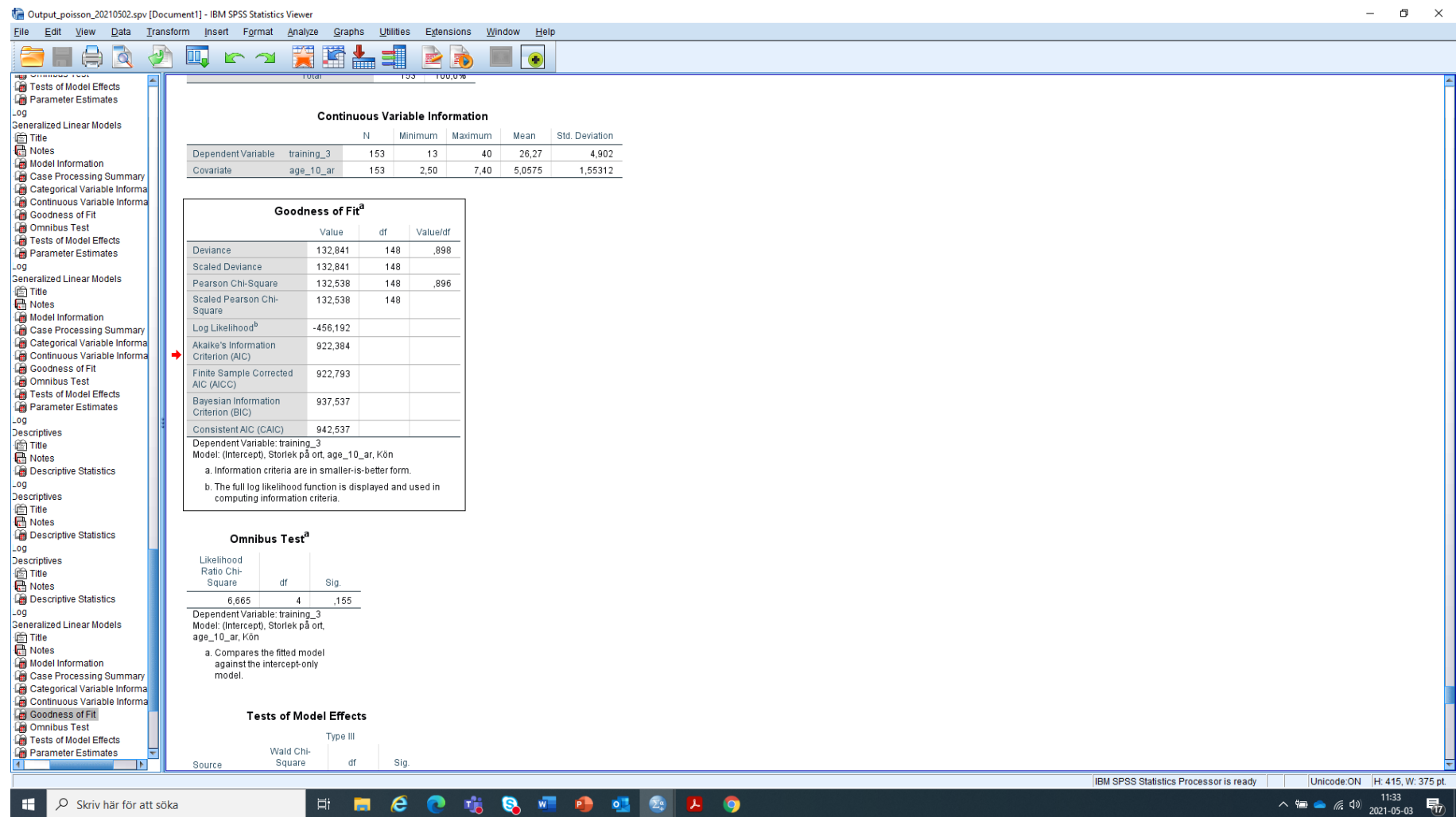
- Den beroende variabeln är ett positivt heltal – antal gymbesök
- Den beroende variabeln följer poissonfördelningen
 - Medelvärde $\approx$ variansen

n=156	medelvärde	varians
Antal träningstillfällen	26,12	24,77

- Test: Kolmogorov-smirnov test
 - $\alpha = 0,05$ resultat p-värde 0.979

Modell

- Beroende variabel: antal gymbesök
- Oberoende variabler:
 - Storlek bostadsort i 3 grupper
<1000 invånare
1000-15 000 invånare
> 15 000 invånare
 - Kön
 - Ålder i 10-tals år



Output_poisson_20210502.spv [Document1] - IBM SPSS Statistics Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help

b. The full log likelihood function is displayed and used in computing information criteria.

Omnibus Test^a

Likelihood Ratio Chi-Square	df	Sig.
6,665	4	,155

Dependent Variable: training_3
Model: (Intercept), Storlek på ort, age_10_ar, Køn

a. Compares the fitted model against the intercept-only model.

Tests of Model Effects

Source	Wald Chi-Square	df	Sig.
(Intercept)	3185,995	1	,000
Storlek på ort	2,489	2	,288
age_10_ar	4,096	1	,043
Køn	,348	1	,555

Dependent Variable: training_3
Model: (Intercept), Storlek på ort, age_10_ar, Køn

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	3,354	,0601	3,236	3,471	3113,592	1	,000	28,603	25,425	32,179
[Storlek på ort=3]	,078	,0498	-,019	,176	2,482	1	,115	1,082	,981	1,192
[Storlek på ort=2]	,017	,0583	-,097	,131	,088	1	,767	1,017	,908	1,140
[Storlek på ort=1]	0 ^a	.	.	.	.	.	.	1	.	.
age_10_ar	-,021	,0104	-,042	-,001	4,096	1	,043	,979	,959	,999
[Køn=2]	,019	,0327	-,045	,083	,348	1	,555	1,019	,956	1,087
[Køn=1]	0 ^a	.	.	.	.	.	.	1	.	.
(Scale)	1 ^b	.	.	.	.	.	.	.	.	.

Dependent Variable: training_3
Model: (Intercept), Storlek på ort, age_10_ar, Køn

a. Set to zero because this parameter is redundant.
b. Fixed at the displayed value.

IBM SPSS Statistics Processor is ready Unicode:ON H: 145, W: 216 pt. 11:19 2021-05-03

Hur lämplig är den här modellen?

Goodness of Fit^a

	Value	df	Value/df
Deviance	132,841	148	,898
Scaled Deviance	132,841	148	
Pearson Chi-Square	132,538	148	,896
Scaled Pearson Chi-Square	132,538	148	
Log Likelihood ^b	-456,192		
Akaike's Information Criterion (AIC)	922,384		
Finite Sample Corrected AIC (AICC)	922,793		
Bayesian Information Criterion (BIC)	937,537		
Consistent AIC (CAIC)	942,537		

- Deviance – hur mycket modellen avviker från poissonfördelningen
- Måttet som används är deviance/ antal frihetsgrader (fg)
- Målvärdet = 1
- Overdispersion > 1
- Underdispersion < 1

Over all effects

Tests of Model Effects

Source	Wald Chi-Square	Type III	
		df	Sig.
(Intercept)	3185,995	1	,000
Storlek på ort	2,489	2	,288
age_10_ar	4,096	1	,043
Kön	,348	1	,555

Resultat poissonregression

Parameter	Parameter Estimates			
	Hypothesis Test	Exp(B)	95% Wald Confidence Interval for Exp(B)	
	Sig.		Lower	Upper
(Intercept)	,000	28,603	25,425	32,179
[Storlek på ort= <1000]	,115	1,082	,981	1,192
[Storlek på ort=1000- 15 000]	,767	1,017	,908	1,140
[Storlek på ort >15 000]	.	1	.	.
age_10_ar	,043	,979	,959	,999
[Kön=Kvinna]	,555	1,019	,956	1,087
[Kön=Man]	.	1	.	.
(Scale)				

Bättre än en nollmodell?

Omnibus Test ^a		
Likelihood Ratio Chi-Square	df	Sig.
6,665	4	,155

Nej!

Dependent Variable: training_3
Model: (Intercept), Storlek på ort,
age_10_ar, Kön^a

a. Compares the fitted model
against the intercept-only model.

Hantera olika tidsintervall

- Vad gör man om tidsintervallen är väldigt olika långa?
 - Exempel: antal besök sedan diagnos.
 - Tiden varierar mycket mellan studiedeltagarna
 - Kommer skillnaden vi eventuellt ser endast bero på spridningen i tid?
- Lösning: Offset variable
 - Kan t.ex. vara tid från diagnos

Exempel

I en större hälsoundersökning frågade man bland annat "Under de senaste 30 dagarna, hur ofta har du upplevt psykisk ohälsa?"

Beroende variabel: antal dagar med psykisk ohälsa

Oberoende variabler:

Race (1= white / 0= non-white)

Ålder

Utbildning

Fysisk hälsa – ju högre score desto bättre hälsa

Resultat poissonmodell – poor mental health during 30 days

Omnibus test - signifikant skilt från 0

Variabel	β	P-värde
Intercept	3.540	<0.001
White	0.202	<0.001
Education	0.124	<0.001
Condition of physical health	-0.781	<0.001
Age	-0.012	<0.001

Resultat poissonmodell –poor mental health during 30 days

Goddness of fit

Deviance	DF	Deviance/df
5567,898	729	7,636

Overdispersion!

	Mdelvärde	varians
Poor mental health	3,9672	56,511

→ Negative binomial regression

Resultat Negative binomial regression

Goodness of fit

Deviance	df	Deviance/df
1512,193	729	2,074

Omnibus test = signifikant skilt från 0

Resultat - Negative binomial regression

Omnibus test - signifikant skilt från 0

Variabel	β	P-värde
Intercept	3.844	<0.001
White	0.045	0.668
Education	0.124	0.001
Condition of physical health	-0.798	<0.001
Age	-0.014	<0.001

Negative binomial parameter = 1

Inställningar negative binomial regression

Generalized Linear Models

Type of Model | Response | Predictors | Model | Estimation | Statistics | EM Means | Save | Export

Choose one of the model types listed below or specify a custom combination of distribution and link function.

Scale Response

- ☐ Linear
- ☐ Gamma with log link

Counts

- ☐ Poisson loglinear
- ☐ Negative binomial with log link

Mixture

- ☐ Tweedie with log link
- ☐ Tweedie with identity link

Custom

- ☒ Custom

Ordinal Response

- ☐ Ordinal logistic
- ☐ Ordinal probit

Binary Response or Events/Trials Data

- ☐ Binary logistic
- ☐ Binary probit
- ☐ Interval censored survival

Distribution: **Negative binomial** Link function: **Log**

Parameter

- ☐ Specify value
Value: 1
- ☒ Estimate value

Power:

OK Paste Reset Cancel Help

Parameter Estimates

		95% Wald Confidence Interval		Hypothesis Test			95% Wald Confidence Interval for Exp(B)			
	R	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.	Exp(B)	Lower	Upper

Resultat Negative binomial regression – justerad

Goodness of fit

Deviance	df	Deviance/df
634,308	728	0,871

Omnibus test = signifikant skilt från 0

Resultat - Negative binomial regression justerat översdiperson

Omnibus test - signifikant skilt från 0

Variabel	β	P-värde
Intercept	3.897	<0.001
White	-0.014	0.939
Education	0.111	0.086
Condition of physical health	-0.799	<0.001
Age	-0.015	0.002

Negative binomial parameter 3,942

Resultat tidigare Poissonregression

Variabel	β	P-värde
Intercept	3.540	<0.001
White	0.202	<0.001
Education	0.124	<0.001
Condition of physical health	-0.781	<0.001
Age	-0.012	<0.001

Jämförelse regressionsmodeller

	Poisson regression	Linjär regression	Logistisk regression	Cox regression
Utfall	Antal events	Kontinuerligt t.ex. längd i centimeter	Binärt t.ex. sjuk/ej sjuk	Tid till event
effektmått	Incidence rate ratio (IRR)	linear coefficient	Odds ratio (OR)	Hazard ratio (HR)
Antagande	Medelvärde = varians	Normalfördelade residualer	Binary outcome	Constant hazard

Poissonregression kan även användas till

- Överlevnadsanalys
 - För att analysera "hazards" uppdelat i olika tidsperioder
 - Modellera absoluta risker
 - Modeller med beroende
- Logistisk regression
 - För att beräkna oddskvoter vid hög prevalens av utfallet i kohortstudier eller tvärsnittsstudier

Läs mer

<https://science.nu/lektion/poisson-regression/>

Generalized linear models – An applied approach; Ulf Olsson, Studentlitteratur (2002)

Exempel binomial regression från Mike Crowson

<https://www.youtube.com/watch?v=6FzRcn72n6o>

<https://stats.idre.ucla.edu/r/dae/poisson-regression/>

<https://stats.idre.ucla.edu/spss/dae/negative-binomial-regression/>

Tack så mycket!

- Nästa föredrag:

SPSS-handhavande.

25 maj klockan 12:00 – 13:00

Erling Englund, Region Västernorrland.

Länk: <https://umu.zoom.us/j/61397725099>

<https://www.norrasjukvardsregionforbundet.se/forskning/forum-norr-for-klinisk-forskning/utbildning/>