Université Panthéon-Assas

Master 2

INGÉNIERIE STATISTIQUE ET FINANCIÈRE

Insurance Econometrics
Regression for Count Data

Joseph Lanfranchi, 2021-22

The effect of insurance contracts on health consumption

- In this last lecture on modelling insurance behaviour, our goal is to explain how access to private or public health insurance induces the patient to increase their health care demand.

- Indeed, risk pooling provided by public or private insurance contract is likely to create moral hazard from the insured party.

- We make a distinction between ex ante and ex post moral hazard. This last term depicts a situation where people who hold health insurance consume more health services than would be optimal. This name has been chosen because the behavior occurs *after* the loss associated with the risk has occurred.

The effect of insurance contracts on health consumption

- This effect arises because insurance companies pay for treatment rather than indemnifying the patient. As it is difficult to observe the necessary health service, patient can chose to over consume.

- On the contrary, the term ex ante moral hazard refers to another type of asymmetric information situation where people chose to underprovide effort for protecting themselves or an asset after signing an insurance contract with an insurance company. This last behavior is denoted as ex-ante moral hazard.

- Moral hazard may not be the only explanation for the positive relationship between health demand and insurance. The causality may be reversed with more fragile individuals choosing to insure themselves more extensively.

The effect of insurance contracts on health consumption

- Health demand studies model data on the number of times that individuals consume a health service, such as visits to a doctor or days in the hospital in a given period.

- In this chapter we will try to explain if the frequency of visits to general practitioners or non medic depends on the nature of the insurance of the patient.

- In this context the dependent or response variable of interest is a nonnegative integer or count that we wish to explain or analyze in terms of a set of regressors. Unlike the classical regression model, the response variable is discrete, with a distribution that places probability mass at nonnegative integer values only.

- For that matter, we will examine how Sas helps to run Poisson regression and negative binomial regression, which are two methods that are appropriate for dependent variables that have only non-negative integer values: 0, 1, 2, 3, etc. Usually these numbers represent counts of something, like number of people in an organization, number of episodes of sickness absenteeism, or number of arrests in the past year.

The effect of insurance contracts on health consumption

- For years, people analyzed count data by ordinary linear regression and, for many applications, that method was adequate for the task.

- However, Poisson and negative binomial regression have the advantage of being precisely tailored to the discrete, often highly skewed distribution of the dependent variable.

- On the downside, Poisson regression has the disadvantage of being susceptible to problems of overdispersion that do not affect ordinary linear regression.

- Overdispersion, discussed in detail later, can produce severe underestimates of standard errors and overestimates of test statistics. While there are some simple corrections for overdispersion, negative binomial regression is generally the preferred method whenever there is evidence for overdispersion.

Structure of the database

- The data are a cross-section sample from the U.S. Medical Expenditure Panel Survey for 2003. The model will use a sample of the Medicare population aged 65 and higher.
- Medicare is a national health insurance program in the United States, begun in 1966 under the Social Security Administration. In general, all persons 65 years of age or older who have been residents of the United States for at least five years are eligible for this program.
- In this database, individual information is reported about individual characteristics – age, gender, years of education, if they are black or Hispanic –, about medical consumption – annual number of doctor visits, annual number of visits to health professional, but not doctor, number of chronic conditions, presence of activity limitation –, and insurance access – public Medicaid insurance, private insurance, employer provided private insurance –.
- Medicaid is a federal and state assistance program that helps with medical costs for some people with limited income and resources. Medicaid also offers benefits not normally covered by Medicare, including nursing home care and personal care services.

Structure of the database

- The interesting characteristics of count data distribution is the shape of this distribution and their first moments.
- The distribution of the variable number of doctor visits has a very long right tail. 22% of the observations exceed 10, and more than 99% of the values are under 40. The proportion of zeros is quite high with 10,9%.
- It should be noted that this percentage of zeros is relatively low for data about doctor visits, probably because the data pertain to the elderly population.
- Samples of younger, and usually healthier population often have as many as 90% zero observations for some health outcomes.
- The distribution of the variable of health professional visits contains a much higher proportion of zero visits (53%).

doctor visits

docvis	Fréquence	Pourcentage	Fréquence cumulée	Pctage. cumulé
0	401	10.91	401	10.91
1	314	8.54	715	19.45
2	358	9.74	1073	29.18
3	334	9.08	1407	38.26
4	339	9.22	1746	47.48
5	266	7.23	2012	54.72
6	231	6.28	2243	61.00
7	202	5.49	2445	66.49
8	179	4.87	2624	71.36
9	154	4.19	2778	75.55
10	108	2.94	2886	78.49
11	127	3.45	3013	81.94
12	89	2.42	3102	84.36
13	85	2.31	3187	86.67
14	81	2.20	3268	88.88
15	70	1.90	3338	90.78
16	51	1.39	3389	92.17
17	43	1.17	3432	93.34
18	33	0.90	3465	94.23
19	27	0.73	3492	94.97
20	26	0.71	3518	95.68
21	19	0.52	3537	96.19
22	21	0.57	3558	96.76
23	17	0.46	3575	97.23
24	15	0.41	3590	97.63
25	6	0.16	3596	97.80
26	5	0.14	3601	97.93
27	11	0.30	3612	98.23
28	4	0.11	3616	98.34
29	6	0.16	3622	98.50
30	8	0.22	3630	98.72
31	2	0.05	3632	98.78
32	6	0.16	3638	98.94
33	3	0.08	3641	99.02

#Visits to health professional, but not doctor

nonphysician	Fréquence	Pourcentage	Fréquence cumulée	Pctage. cumulé
0	1949	53.01	1949	53.01
1	587	15.96	2536	68.97
2	289	7.86	2825	76.83
3	182	4.95	3007	81.78
4	128	3.48	3135	85.26
5	88	2.39	3223	87.65
6	69	1.88	3292	89.53
7	46	1.25	3338	90.78
8	40	1.09	3378	91.87
9	34	0.92	3412	92.79
10	32	0.87	3444	93.66
11	20	0.54	3464	94.21
12	27	0.73	3491	94.94
13	17	0.46	3508	95.40
14	11	0.30	3519	95.70
15	13	0.35	3532	96.06
16	17	0.46	3549	96.52
17	10	0.27	3559	96.79
18	10	0.27	3569	97.06
19	10	0.27	3579	97.33
20	6	0.16	3585	97.50
21	9	0.24	3594	97.74
22	7	0.19	3601	97.93
23	5	0.14	3606	98.07
24	9	0.24	3615	98.31
25	4	0.11	3619	98.42
26	6	0.16	3625	98.59
27	2	0.05	3627	98.64
28	3	0.08	3630	98.72
31	2	0.05	3632	98.78
32	3	0.08	3635	98.86
33	1	0.03	3636	98.88
34	3	0.08	3639	98.97
35	3	0.08	3642	99.05

Procédure MEANS

Variable	Libellé	N	Moyenne	Ecart-type	Minimum	Maximum
actlim	=1 if activity limitation	3677	0.3331520	0.4714045	0	1.0000000
age	Age	3677	74.2447648	6.3766378	65.0000000	90.0000000
bh	=1 if black or Hispanic	3677	0.2561871	0.4365857	0	1.0000000
docvis	# doctor visits	3677	6.8226815	7.3949367	0	144.0000000
educyr	Years of education	3677	11.1803100	3.8276759	0	17.0000000
female	=1 if female	3677	0.6010335	0.4897525	0	1.0000000
insured	=1 if has private supplementary insurance	3677	0.4966005	0.5000564	0	1.0000000
medicaid	=1 if has Medicaid public insurance	3677	0.1667120	0.3727692	0	1.0000000
nonphysician	#Visits to health professional, but not doctor	3677	2.7166168	7.8748493	0	187.0000000
offer	=1 if employer offers insurance	3677	0.0339951	0.1812412	0	1.0000000
phylim	=1 if physical limitation	3677	0.4666848	0.4989567	0	1.0000000
totchr	# chronic conditions	3677	1.8433506	1.3500262	0	8.0000000

Variable : docvis (# doctor visits)

Mesures statistiques de base

Tendance centrale		Variabilité	
Moyenne	6.822682	Ecart-type	7.39494
Médiane	5.000000	Variance	54.68509
Mode	0.000000	Intervalle	144.00000
		Ecart interquartile	7.00000

nonphysician (#Visits to health professional, but not doctor)

Mesures statistiques de base

Tendance centrale		Variabilité	
Moyenne	2.716617	Ecart-type	7.87485
Médiane	0.000000	Variance	62.01325
Mode	0.000000	Intervalle	187.00000
		Ecart interquartile	2.00000

Count data model: The Poisson regression model

- The Poisson is the starting point for count data analysis, though it is often inadequate. This implies a Poisson distribution for the number of occurrences of the event, with density, or more formally probability mass function :

$$\Pr(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}, y = 0,1,2, \dots$$

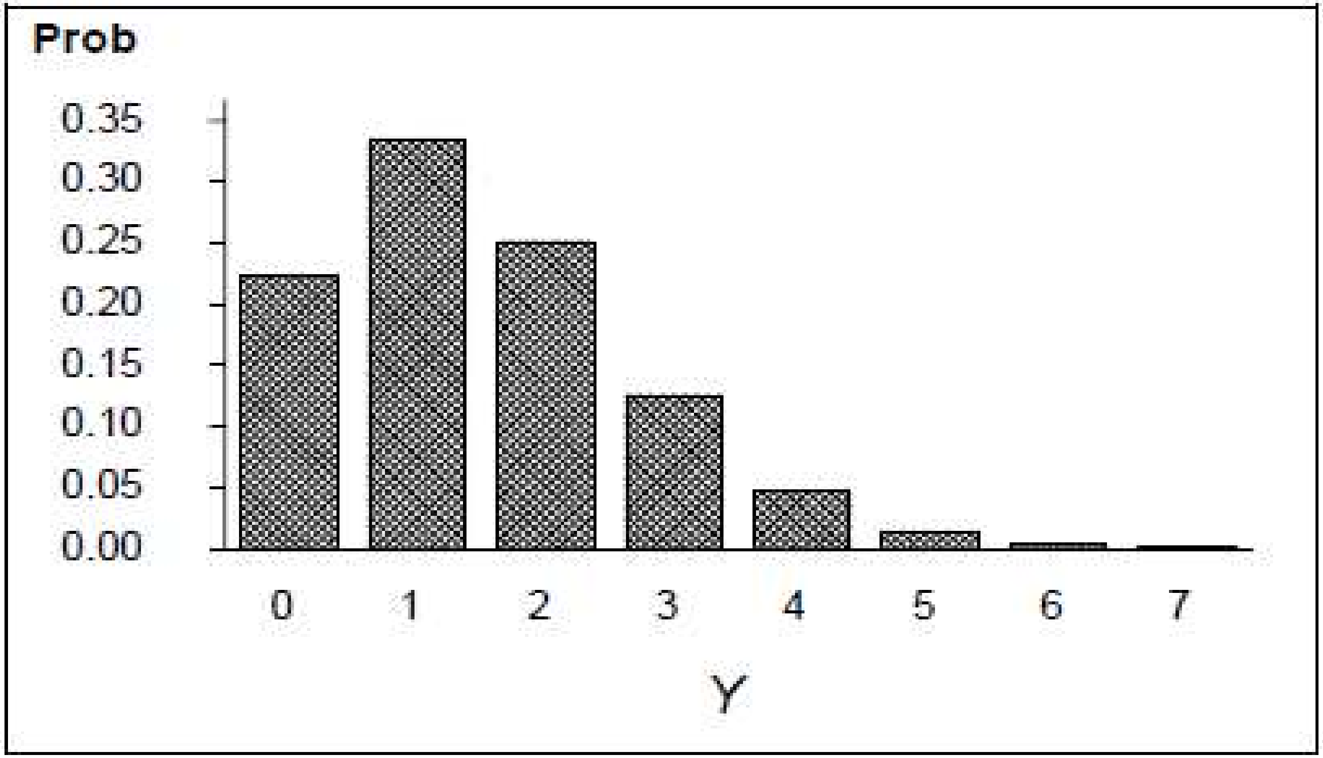
where μ is an intensity or rate parameter.

- The first two moments of this distribution are:

$$\begin{cases} E(Y) = \mu \\ Var(Y) = \mu \end{cases}$$

- So the parameter of intensity is equal to the expected value of the count variable. Furthermore, this set of equalities shows the well-known equidispersion (equality of mean and variance) property of the Poisson distribution.

- As μ gets larger, the mode of the distribution moves away from 0 and the distribution looks more and more like a normal distribution. For an unitary value of the intensity parameter, the probability that *Y=0* equals 0,368. For a value of 5, this probability falls to 0,0067.



This distribution is the theoretical distribution when $\mu = 1{,}5$

The Poisson regression model (2)

- We need to specify how the parameter μ depends on the explanatory variables. First, we write μ_i with a subscript i to allow the parameter to vary across individuals: $i = 1, \ldots, n$. Then the Poisson regression model is derived from the Poisson distribution by parameterizing the relation between the mean parameter and covariates (regressors) x.
- A first requirement comes from the fact that μ_i cannot be less than 0. Hence, it is standard to let μ_i be a loglinear function of the x variables. This ensures that μ_i will be greater than 0 for any values of the x's or the β's:

$$\log \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \quad (1)$$

- This model specification can also be written as:

$$\mu_i = \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right)$$

- That's all there is to say about the model. By the property of equality of mean and variance of the count variable *Y*, we can assess that the Poisson regression is intrinsically heteroskedastic:

$$V(Y_i|x_i) = \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right)$$

The Poisson regression model (3)

- We'll estimate the model by maximum likelihood. This is accomplished with PROC GENMOD.

- Note that the model does not say that the marginal distribution of y will necessarily be Poisson. Instead, y has a Poisson distribution conditional on the values of the explanatory variables. If the x variables have large coefficients and large variances, the marginal distribution of y may look very different from a Poisson distribution.

- In that case, it may be necessary to change the hypothesis about the random process generating the observations.

- To fit a Poisson regression, we use the DIST=POISSON option, which can be abbreviated D=P. When a Poisson regression is requested, the loglinear model in equation (1) above is the default.

The Poisson regression model (4)

- In the output we can see that all variables are significant to explain the number of doctor visits. Notice that the two private and social insurance have a positive effect on the extent of doctors visits. Among the other determinants of doctor visits, activity limitation and the number of chronic conditions have also positive effects, like manhood.

- For the health professional visits, individuals privately insured are still visiting doctors more often but that is not the case of Medicaid utilizers. This social program does not pay for giving access to non doctor health care. Hence, the Medicaid dummy has to be interpreted as an indicator of having limited assets. Then, poor elderly people are not sufficiently wealthy to finance health professional visit.

- Another change in the results: women are more likely to demand this type of visit than men.

Variable dépendante	docvis	# doctor visits

Analyse des paramètres estimés du maximum de vraisemblance

Paramètre	DDL	Estimation	Erreur type	Intervalle de confiance de Wald à95%		Khi-2 de Wald	Pr > khi-2
Intercept	1	-10.1210	0.9719	-12.0259	-8.2160	108.44	<.0001
private	1	0.1222	0.0144	0.0939	0.1505	71.59	<.0001
medicaid	1	0.1372	0.0193	0.0993	0.1750	50.34	<.0001
age	1	0.2962	0.0260	0.2453	0.3470	130.19	<.0001
age2	1	-0.0020	0.0002	-0.0023	-0.0016	128.30	<.0001
educyr	1	0.0249	0.0019	0.0211	0.0287	164.57	<.0001
female	1	-0.0484	0.0131	-0.0741	-0.0226	13.56	0.0002
bh	1	-0.1596	0.0170	-0.1928	-0.1263	88.52	<.0001
actlim	1	0.1873	0.0146	0.1587	0.2159	165.18	<.0001
totchr	1	0.2487	0.0047	0.2396	0.2578	2855.92	<.0001
Echelle	0	1.0000	0.0000	1.0000	1.0000		

Variable dépendante	nonphysician	#Visits to health professional, but not doctor

Analyse des paramètres estimés du maximum de vraisemblance

Paramètre	DDL	Estimation	Erreur type	Intervalle de confiance de Wald à95%		Khi-2 de Wald	Pr > khi-2
Intercept	1	-19.4884	1.5689	-22.5634	-16.4135	154.31	<.0001
private	1	0.1548	0.0222	0.1112	0.1984	48.47	<.0001
medicaid	1	-0.3664	0.0375	-0.4400	-0.2929	95.25	<.0001
age	1	0.5070	0.0419	0.4249	0.5891	146.57	<.0001
age2	1	-0.0034	0.0003	-0.0039	-0.0028	145.75	<.0001
educyr	1	0.0787	0.0034	0.0721	0.0853	543.78	<.0001
female	1	0.0962	0.0209	0.0553	0.1371	21.25	<.0001
bh	1	-0.4519	0.0308	-0.5123	-0.3914	214.57	<.0001
actlim	1	0.3443	0.0229	0.2994	0.3892	226.00	<.0001
totchr	1	0.2060	0.0075	0.1914	0.2207	759.53	<.0001
Echelle	0	1.0000	0.0000	1.0000	1.0000		

Meaning of estimated coefficients in the Poisson model

INGÉNIERIE STATISTIQUE & FINANCIÈRE
UNIVERSITE PANTHEON ASSAS

- Because the dependent variable is logged, we can interpret the coefficients much like logistic regression coefficients.
- According to the count model:

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}),$$

the coefficient β_1 evaluates the percentage change in the number of visits of a one unit change in the variable x_1.

- In our regression results, this means that a private insurance contract is associated with a 12,2% increase in the number of doctor visits and with a 15,5% increase in the health professional visits. However, being insured by the program Medicaid is associated with an increase of 13,7% of doctor visits but with a decrease of 37% of other health visits.

UNIVERSITÉ PARIS II
PANTHÉON-ASSAS

Meaning of estimated coefficients in the Poisson model

- Among the variables measuring the health state, suffering from physical limitation raises the number of doctor visits by 18,7%, while every supplementary chronic disease implies a 24,9% increase of these visits.
- Women have 4,8% less doctor visits than men but 9,6% more health professional visits.
- For the age, we can say that the effect is increasing then decreasing on the number of doctor visits with a maximum at 74,05 years old (-29.62/2*-0.20).

Overdispersion of observations

- Overdispersion can be detected using the two goodness-of-fit chi-squares, the deviance and the Pearson chi-square. In the Poisson model, the variance of the dependent variable should be equal to its mean. In fact, the variance is often much higher than that. Hence, we must take into account the overdispersion of observations.

- Note that in our results for the number of doctor visits, the deviance is 5 times as large as the number of degrees of freedom. This large ratio of deviance to degrees of freedom does suggest an overdispersion problem with the model.

- To recognize an overdispersion problem, we take the ratio of the goodness-of-fit chi-square to its degrees of freedom, and compare the result with one. When those ratios are largely superior to unity, overdispersion of observations is very likely.

Variable dépendante	docvis		# doctor visits
Critères d'évaluation de l'adéquation			
Critère	DDL	Valeur	Valeur/DDL
Ecart	3667	18288.4460	4.9873
Déviance normalisée	3667	18288.4460	4.9873
Khi2 de Pearson	3667	23232.8352	6.3357
Pearson normalisé X2	3667	23232.8352	6.3357

Variable dépendante	nonphysician		#Visits to health professional, but not doctor
Critère	DDL	Valeur	Valeur/DDL
Ecart	3667	27226.7059	7.4248
Déviance normalisée	3667	27226.7059	7.4248
Khi2 de Pearson	3667	79163.9009	21.5882
Pearson normalisé X2	3667	79163.9009	21.5882

Overdispersion of observations

- It's not always appropriate to calculate a p-value for this statistic because the predicted values of Y is quite small for many of the elderly people.
- In such a case, when predicted values are small, the deviance is not well approximated by a chi-square distribution.
- Large overdispersion leads to grossly deflated standard errors and grossly inflated t-statistics in the usual ML output, and hence it is important to correct for that problem if you want to use statistical tests for the significance of the estimated coefficients.
- Hence, in our results, we can believe in the value of these coefficients but we have seen that their t statistics are very high.

Overdispersion of observations (II)

- What can we do about the problem of overdispersion?
- First, we have to remember that provided the conditional mean is correctly specified, the Poisson MLE is still consistent.
- Hence, in a first step, it's a simple matter to correct the standard errors and chi-squares.
- It is then necessary to determine a dispersion parameter and increase the variance-covariance matrix of estimated parameters with the value of this parameter.

Overdispersion of observations (III)

- The ratio of the goodness-of-fit chi-square to its degrees of freedom is estimated using the deviance or the Pearson chi-square. These dispersion parameters are fairly close, but the theory of MLE suggests the use of the Pearson chi-square.

- Method: take the ratio of the goodness-of-fit chi-square to its degrees of freedom, and call the result C. Then, divide the chi-square statistic for each coefficient by C. Finally, multiply the standard error of each coefficient by the square root of C.

- In Sas, the corrections just described can be automatically invoked by putting either SCALE=P (for Pearson) or SCALE=D (for deviance) as options in the MODEL statement of the proc GENMOD.

- The only variable losing significance at 5% level is the individual gender, that does not explain the number of visits to any type of health practitioner.

Critère	DDL	Valeur	Valeur/DDL
Ecart	3667	18288.4460	4.9873
Déviance normalisée	3667	3667.0000	1.0000
Khi2 de Pearson	3667	23232.8352	6.3357
Pearson normalisé X2	3667	4658.3951	1.2704

Critère	DDL	Valeur	Valeur/DDL
Ecart	3667	27226.7059	7.4248
Déviance normalisée	3667	3667.0000	1.0000
Khi2 de Pearson	3667	79163.9009	21.5882
Pearson normalisé X2	3667	10662.1060	2.9076

Analyse des paramètres estimés du maximum de vraisemblance							
Paramètre	DDL	Estimation	Erreur type	Intervalle de confiance de Wald à95%		Khi-2 de Wald	Pr > khi-2
Intercept	1	-10.1210	2.1705	-14.3752	-5.8668	21.74	<.0001
private	1	0.1222	0.0322	0.0590	0.1854	14.35	0.0002
medicaid	1	0.1372	0.0432	0.0525	0.2218	10.09	0.0015
age	1	0.2962	0.0580	0.1825	0.4098	26.11	<.0001
age2	1	-0.0020	0.0004	-0.0027	-0.0012	25.73	<.0001
educyr	1	0.0249	0.0043	0.0164	0.0334	33.00	<.0001
female	1	-0.0484	0.0293	-0.1059	0.0091	2.72	0.0991
bh	1	-0.1596	0.0379	-0.2338	-0.0853	17.75	<.0001
actlim	1	0.1873	0.0325	0.1235	0.2511	33.12	<.0001
totchr	1	0.2487	0.0104	0.2283	0.2690	572.64	<.0001
Echelle	0	2.2332	0.0000	2.2332	2.2332		

Analyse des paramètres estimés du maximum de vraisemblance							
Paramètre	DDL	Estimation	Erreur type	Intervalle de confiance de Wald à 95%		Khi-2 de Wald	Pr > khi-2
Intercept	1	-19.4884	4.2749	-27.8672	-11.1097	20.78	<.0001
private	1	0.1548	0.0606	0.0361	0.2736	6.53	0.0106
medicaid	1	-0.3664	0.1023	-0.5670	-0.1659	12.83	0.0003
age	1	0.5070	0.1141	0.2834	0.7307	19.74	<.0001
age2	1	-0.0034	0.0008	-0.0048	-0.0019	19.63	<.0001
educyr	1	0.0787	0.0092	0.0607	0.0967	73.24	<.0001
female	1	0.0962	0.0569	-0.0153	0.2077	2.86	0.0907
bh	1	-0.4519	0.0841	-0.6166	-0.2871	28.90	<.0001
actlim	1	0.3443	0.0624	0.2220	0.4666	30.44	<.0001
totchr	1	0.2060	0.0204	0.1661	0.2460	102.30	<.0001

Some proximity with the OLS

- Ordinary linear regression is not susceptible to the problem of overdispersion because it automatically estimates a scale parameter that is used in calculating standard errors and test statistics. The scale parameter for a linear regression is just the estimated standard deviation of the disturbance term, sometimes called the root mean squared error.
- To illustrate this point, we can use ordinary least squares (OLS) to regress the two models and the results, while not identical to those in previous output, are quite close, as we can see from the age effect (max at 77,35).
- In general, Poisson regression with a correction for overdispersion is better than ordinary least squares but OLS may be better than Poisson regression without the overdispersion correction.

Analyse des paramètres estimés du maximum de vraisemblance							
Paramètre	DDL	Estimation	Erreur type	Intervalle de confiance de Wald à95%		Khi-2 de Wald	Pr > khi-2
Intercept	1	-11.2128	2.3180	-15.7560	-6.6695	23.40	<.0001
private	1	0.1837	0.0349	0.1153	0.2520	27.75	<.0001
medicaid	1	0.1224	0.0484	0.0276	0.2172	6.40	0.0114
age	1	0.3101	0.0620	0.1886	0.4317	25.01	<.0001
age2	1	-0.0020	0.0004	-0.0028	-0.0012	24.23	<.0001
educyr	1	0.0219	0.0047	0.0128	0.0311	22.22	<.0001
female	1	0.0327	0.0321	-0.0302	0.0956	1.04	0.3086
bh	1	-0.2014	0.0402	-0.2802	-0.1225	25.08	<.0001
actlim	1	0.1435	0.0372	0.0707	0.2164	14.91	0.0001
totchr	1	0.3329	0.0124	0.3086	0.3573	717.99	<.0001
Echelle	1	0.9405	0.0110	0.9192	0.9622		

Analyse des paramètres estimés du maximum de vraisemblance							
Paramètre	DDL	Estimation	Erreur type	Intervalle de confiance de Wald à95%		Khi-2 de Wald	Pr > khi-2
Intercept	1	-12.3955	2.7055	-17.6981	-7.0929	20.99	<.0001
private	1	0.1342	0.0407	0.0545	0.2140	10.88	0.0010
medicaid	1	-0.2167	0.0565	-0.3274	-0.1060	14.72	0.0001
age	1	0.3189	0.0724	0.1771	0.4608	19.42	<.0001
age2	1	-0.0021	0.0005	-0.0031	-0.0012	19.80	<.0001
educyr	1	0.0474	0.0054	0.0368	0.0580	76.16	<.0001
female	1	0.1311	0.0375	0.0576	0.2045	12.24	0.0005
bh	1	-0.3710	0.0469	-0.4630	-0.2790	62.49	<.0001
actlim	1	0.0634	0.0434	-0.0217	0.1484	2.13	0.1441
totchr	1	0.1578	0.0145	0.1293	0.1862	118.38	<.0001
Echelle	1	1.0977	0.0128	1.0728	1.1230		

Adjustment of overdispersion with Negative Binomial Regression

- Efficient estimates may be produced by a method known as negative binomial regression that has become increasingly popular for count data.
- The negative binomial model is a generalization of the Poisson model. We modify equation () to include a disturbance term, which accounts for the overdispersion:

$$\log \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \sigma \varepsilon_i$$

- The dependent variable y_i has a Poisson distribution with expected value μ_i, ***conditional on*** ε_i. Finally, we assume that $\exp(\varepsilon_i)$has a standard gamma distribution and It follows that ***the unconditional distribution of*** y_i ***is a negative binomial distribution.***
- The negative binomial regression model may be efficiently estimated by maximum likelihood. In PROC GENMOD, this is accomplished simply by using the option D=NB on the MODEL statement.
- Now the deviance is actually a little bit superior to the degrees of freedom, indicating a reasonable fit and a good correction of overdispersion. If the dispersion parameter were 0, we would be back to the Poisson model.

UNIVERSITÉ PARIS II
PANTHÉON-ASSAS

Negative Binomial Regression

INGÉNIERIE STATISTIQUE FINANCIÈRE
UNIVERSITE PANTHEON ASSAS

- You can get a test for whether the dispersion parameter is 0 by putting the option NOSCALE on the MODEL statement. This constrains the dispersion parameter to be 0 and reports a Lagrange multiplier test for that constraint.
- For this example, we get a chisquare of 1715,6 with one degree of freedom, which is highly significant. So we reject the simpler Poisson model in favor of the more complicated negative binomial model.
- In this new model, the precision of the estimated parameters has improved with the p-values of the significant variables noticeably lower.
- The logic of the results for the model explaining the number of visits to health professionals is quite similar with a low ratio deviance / number of degrees of freedom (less than 1).
- This globally confirms the superiority of the negative binomial model.

UNIVERSITÉ PARIS II
PANTHÉON-ASSAS

Critères d'évaluation de l'adéquation			
Critère	DDL	Valeur	Valeur/DDL
Ecart	3667	4196.3454	1.1444
Déviance normalisée	3667	4196.3454	1.1444
Khi2 de Pearson	3667	4688.4230	1.2785
Pearson normalisé X2	3667	4688.4230	1.2785
Critère	DDL	Valeur	Valeur/DDL
Ecart	3667	18288.4460	4.9873
Déviance normalisée	3667	18288.4460	4.9873
Khi2 de Pearson	3667	23232.8352	6.3357
Pearson normalisé X2	3667	23232.8352	6.3357

Statistiques du multiplicateur de Lagrange			
Paramètre	Khi-2	Pr > khi-2	
Dispersion	1715.6332	<.0001	*
* p-value unilatérale			

Analyse des paramètres estimés du maximum de vraisemblance							
Paramètre	DDL	Estimation	Erreur type	Intervalle de confiance de Wald à95%		Khi-2 de Wald	Pr > khi-2
Intercept	1	-10.1363	2.2437	-14.5339	-5.7386	20.41	<.0001
private	1	0.1491	0.0334	0.0836	0.2146	19.92	<.0001
medicaid	1	0.1470	0.0468	0.0553	0.2387	9.87	0.0017
age	1	0.2941	0.0600	0.1764	0.4118	23.99	<.0001
age2	1	-0.0019	0.0004	-0.0027	-0.0012	23.48	<.0001
educyr	1	0.0231	0.0044	0.0144	0.0317	27.14	<.0001
female	1	-0.0090	0.0306	-0.0690	0.0510	0.09	0.7677
bh	1	-0.1624	0.0399	-0.2407	-0.0841	16.53	<.0001
actlim	1	0.1875	0.0347	0.1195	0.2555	29.23	<.0001
totchr	1	0.2768	0.0121	0.2531	0.3006	520.62	<.0001
Dispersion	1	0.6367	0.0196	0.5994	0.6762		

Zero inflated models of count data

INGÉNIERIE STATISTIQUE FINANCIÈRE
UNIVERSITE PANTHEON ASSAS

- For some applications, the number of individuals with a count of zero may be a large fraction of the sample.
- In the data set just examined, a little more than 10 percent of the patients had zero visits to a doctor. For those patients older than 65, 53 percent never visited any non doctor health professional.
- Poisson regression models often fit poorly when the fraction of zeros is large. This has led to the development of zero-inflated Poisson regression models (ZIP model) which give special treatment to the zero counts.
- The zero-inflated Poisson (ZIP) model is now available in PROC GENMOD, along with a zero-inflated negative binomial model.

UNIVERSITÉ PARIS II
PANTHÉON-ASSAS

Zero inflated models of count data

INGÉNIERIE STATISTIQUE FINANCIÈRE
UNIVERSITE PANTHEON ASSAS

- The goal of such a method is to explain the generation of the observed data with a combination of two models.
- The zero-inflated model supplements the count density $f_2(.)$ with a binary process with density $f_1(.)$. If the binary process takes value 0, with probability $f_1(0)$, then $y = 0$. If the binary process takes value 1, with probability $f_1(0)$, then y takes count values 0,1,2,... from the count density $f_2(\cdot)$.
- This lets zero counts occur in two ways: as a realization of the binary process and as a realization of the count process when the binary random variable takes value 1.
- The first model explains that individuals belong to the zero group. The second model explains the behaviour of individuals with a number of visits that can be superior or equal to 0.

Zero inflated models of count data (2)

- This sort of model is sometimes called a finite mixture model. It can be estimated by maximum likelihood, even though we can't distinguish with certainty whether individuals with counts of zero are in one group or the other.
- In addition to the usual regression coefficients (for the individuals in the regression group), we can get an estimate of the probability that an individual is in the zero group. And we can elaborate the models further by allowing the probability of being in the zero group to be a function of covariates, usually via logistic regression.
- Here is an example of a zero inflated Poisson model explaining the doctor visits, but with no set of explanatory variable for the binary model explaining that individuals belong to the zero group :

- **proc genmod** data=lib.docvisit;
- model nonphysician = private medicaid age age2 educyr female phylim totchr / D=ZIP;
- zeromodel;
- **run**;

Zero inflated models of count data (3)

- The last table of the output is titled « Paramètres estimés par l'analyse du maximum de vraisemblance - Zéro inflation ».
- For the model without explanatory variables, the reported value in the table (0.0962) is the estimated logarithm of the odds to belong to the zero group. Transforming this parameter with the logistic transformation $1/(1 + \exp(-\beta))$, we obtain a value of 0.524.
- According to this model, 52,4% (to compare with the observed 53%) of the sample members are estimated to be in the zero group, which has no chance of experiencing an event (visit to a health professional).
- Since 53 percent of the sample had a count of 0, that means that only 0,6 percent of the sample had counts of 0 but were not in the zero group.

Analyse des paramètres estimés du maximum de vraisemblance							
Paramètre	DDL	Estimation	Erreur type	Intervalle de confiance de Wald à95%		Khi-2 de Wald	Pr > khi-2
Intercept	1	-12.7201	1.6402	-15.9348	-9.5054	60.15	<.0001
private	1	0.0297	0.0225	-0.0144	0.0739	1.74	0.1867
medicaid	1	-0.1113	0.0395	-0.1888	-0.0338	7.92	0.0049
age	1	0.3614	0.0438	0.2755	0.4472	68.13	<.0001
age2	1	-0.0024	0.0003	-0.0029	-0.0018	66.05	<.0001
educyr	1	0.0345	0.0035	0.0275	0.0414	94.62	<.0001
female	1	-0.0179	0.0213	-0.0596	0.0239	0.70	0.4019
bh	1	0.0179	0.0322	-0.0452	0.0810	0.31	0.5776
actlim	1	0.3124	0.0230	0.2674	0.3575	184.77	<.0001
totchr	1	0.1007	0.0077	0.0855	0.1159	169.08	<.0001
Echelle	0	1.0000	0.0000	1.0000	1.0000		

Paramètres estimés par l'analyse du maximum de vraisemblance							
Paramètre	DDL	Estimation	Erreur type	Intervalle de confiance de Wald à 95%		Khi-2 de Wald	Pr > khi-2
Intercept	1	0.0962	0.0335	0.0305	0.1618	8.25	

Zero inflated models of count data (4)

- We can then introduce variables to explain why individuals could belong to the zero group.
- For example here, we have explained this fact with the whole set of explanatory variables of the count model.
- For example, a private insurance contract reduces the probability to report a zero but increases the expected number of visits. Being insured with Medicaid on the contrary increases the probability to be part of the zero group but decreases the expected number of visits for those outside the zero group.
- Another result, being a woman decreases the probability to belong to the zero group but has no effect on the number of visits.
- The higher the number of chronic conditions, the less likely to belong to the zero group and the higher the expected number to health professionals.

Analyse des paramètres estimés du maximum de vraisemblance							
Paramètre	DDL	Estimation	Erreur type	Intervalle de confiance de Wald à 95%		Khi-2 de Wald	Pr > khi-2
Intercept	1	-11.1841	1.6198	-14.3588	-8.0094	47.68	<.0001
private	1	0.0127	0.0223	-0.0311	0.0564	0.32	0.5709
medicaid	1	-0.0874	0.0376	-0.1611	-0.0137	5.40	0.0201
age	1	0.3229	0.0432	0.2381	0.4077	55.75	<.0001
age2	1	-0.0021	0.0003	-0.0027	-0.0016	54.11	<.0001
educyr	1	0.0282	0.0034	0.0216	0.0348	69.55	<.0001
female	1	-0.0364	0.0212	-0.0779	0.0051	2.96	0.0853
phylim	1	0.2734	0.0227	0.2290	0.3179	145.20	<.0001
totchr	1	0.0995	0.0077	0.0843	0.1146	165.47	<.0001
Echelle	0	1.0000	0.0000	1.0000	1.0000		

Paramètres estimés par l'analyse du maximum de vraisemblance							
Paramètre	DDL	Estimation	Erreur type	Intervalle de confiance de Wald à 95%		Khi-2 de Wald	Pr > khi-2
Intercept	1	16.4669	5.2809	6.1166	26.8173	9.72	0.0018
private	1	-0.3704	0.0770	-0.5213	-0.2196	23.16	<.0001
medicaid	1	0.6415	0.1129	0.4203	0.8628	32.29	<.0001
age	1	-0.3847	0.1412	-0.6614	-0.1080	7.43	0.0064
age2	1	0.0026	0.0009	0.0007	0.0044	7.56	0.0060
educyr	1	-0.1179	0.0108	-0.1390	-0.0968	120.00	<.0001
female	1	-0.2424	0.0733	-0.3861	-0.0987	10.93	0.0009
phylim	1	-0.2424	0.0792	-0.3977	-0.0871	9.36	0.0022
totchr	1	-0.2451	0.0287	-0.3014	-0.1889	72.95	<.0001

	Poisson	ZIP	NB	ZINB	ZIP with Expl.	ZINB with expl.
Full Log-vraisemblance	-15011.4	-12427.39	-6724.17	-6724.17	-12225.82	**-6667.44**
BIC	30096.86	24936.88	13530.43	13538.64	24599.42	**13490.88**

The model with the best log likelihood and the weakest BIC statistic is the ZINB model with explanatory variables.