Modelling "Count" Data

We have considered models where the dependent variable is "binary" (0, 1). The Logit & Probit models also apply to cases where the dependent variable is "multinomial".

However, this is for the situation where the dependent variable is qualitative & we code up a y-variable accordingly.

In some cases, we want to model quantitative "count" data, where y takes the values 0, 1, 2, 3, 4 etc.

Why is this a special situation?

Why not just use OLS?

* The variable to be explained is discrete – OLS ignores this information, so OLS cannot be *efficient*.
* If we use MLE, can't assume Normal errors (OLS) – not continuous. So, OLS is mis-specified MLE & will be *inconsistent*, too.

Examples:

* Number of people laid-off
* Number of appliances purchased by household each year.

So, we need a distributional model that allows for the discrete nature of the data. Lots of possibilities!

In many applications, variable is likely to exhibit lots of "0" & "small" values, so want a discrete distribution that allows for this. Then use MLE.

One good & popular such choice is the Poisson Distribution. We have met this in basic MLE examples:

$$Pr.(X = x) = \lambda^{x} e^{-\lambda} / x! \quad \begin{cases} x = 0, 1, 2, 3, \ldots \\ \lambda > 0 \end{cases}$$

$$E(X) = var.(X) = \lambda .$$

In Poisson **Regression** each y_i is assumed to be drawn from a Poisson Dist'n., with parameter λ_i

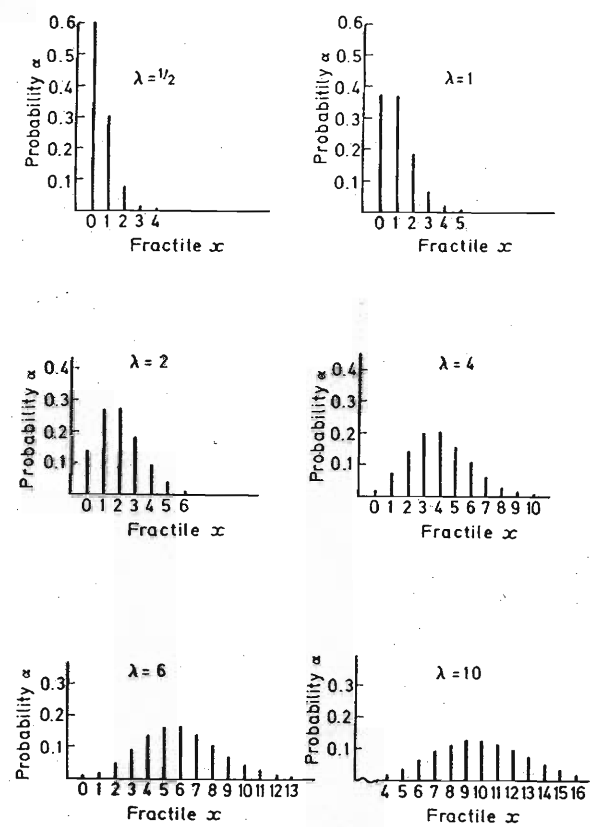


Figure 24.1 Probability function for the Poisson variate P: λ

It would be tempting to then specify:

$$\underset{(1\times1)}{\lambda_i} = \underset{(1\times k)(k\times1)}{\beta' x_i},$$

but recall that $\lambda_i > 0$, as it is the mean of non-negative integers. One way to deal with this is to specify

$$\log(\lambda_i) = \beta' x_i .$$

Then the expected number of events per period is:

$$E(y_i \mid x_i) = \lambda_i = \exp[\beta' x_i].$$

$$(= \text{var.}(y_i \mid x_i))$$

As usual, the partial impact of x_i on y_i is of interest:

$$\frac{\partial E[y_i \mid \underset{\sim}{x_i}]}{\partial \underset{\sim}{x_i}} = \exp[\underset{\sim}{\beta}' x_i]\underset{\sim}{\beta} = \lambda_i \cdot \underset{\sim}{\beta}.$$

(This is a vector of impacts for each element of x_i vector.)

As usual, <u>Maximum Likelihood</u> is the sensible way to estimate the parameters of the Poisson Regression model:

Assuming independence of the y_i's –

$$L = p(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} \lambda_i^{y_i} e^{-\lambda_i}/y_i!$$

$$= (\prod_i \lambda_i^{y_i}) \exp[-\sum_i \lambda_i] / [\prod_i y_i!]$$

$$\log L = \sum_i y_i \log\lambda_i - \sum_i \lambda_i - \sum_i \log y_i!$$

$$= \sum_i y_i \beta' x_i - \sum_i \lambda_i - \sum_i \log y_i!$$

So:

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} y_i x_i - \sum_i \lambda_i x_i = 0$$

$$\left[\lambda_i = \exp(\beta' x_i), \text{ so } \frac{\partial \lambda_i}{\partial \beta} = x_i \exp(\beta' x_i) = \lambda_i x_i \right].$$

The Hessian is:

$$\left[\frac{\partial^2 \log L}{\partial\beta\,\partial\beta'}\right] = -\sum_i \lambda_i x_i x_i'$$

This is Negative-Definite for all x_i & β.

So, we can use the Newton-Raphson algorithm with confidence in this model.

The asymptotic std. errors for the elements of $\tilde{\beta}$ just come from $[\sum_i \tilde{\lambda}_i x_i x_i']^{-1}$, (note the change of sign) where $\tilde{\lambda}_i = \exp[\tilde{\beta}' x_i]$.

The "predicted values" are:

$$\tilde{E}(y_i | x_i) = \tilde{\lambda}_i = \exp[\tilde{\beta}' x_i].$$

If we want to test if there is "no model", then the LRT can be used, as for the Logit & Probit models.

$$\log \tilde{L}_1 = \sum_i y_i \tilde{\beta}' x_i - \sum_i \tilde{\lambda}_i - \sum_i \log y_i!$$

$$\log \tilde{L}_0 = \sum_i y_i \tilde{\beta}_0 - \sum_i \exp(\tilde{\beta}_0) - \sum_i \log y_i!$$

$$= \tilde{\beta}_0 n \bar{y} - n \exp(\tilde{\beta}_0) - \sum_i \log y_i!$$

(where $\tilde{\beta}_0$ is the estimated intercept coeff.).

So:

$$LRT = -2 \log\left(\tilde{L}_0 / \tilde{L}_1\right) = 2[\log \tilde{L}_1 - \log \tilde{L}_0]$$

$$= 2\left[\sum_i y_i \tilde{\beta}' x_i - \sum_i \tilde{\lambda}_i - \tilde{\beta}_0 n \bar{y} + n \exp(\tilde{\beta}_0)\right]$$

$$\xrightarrow{d} \chi^2_{(k-1)}.$$

Goodness-of-Fit:

In the standard regression model, R^2 is defined naturally, because the conditional mean of y is linear. In the Poisson regression model, the usual R^2 does not apply because $E(y_i | x_i)$ is non-linear:

$E(y_i|x_i) = \lambda_i = \exp[\beta' x_i]$.

Moreover, note that the Poisson regression is intrinsically heteroskedastic, & this also complicates the formulation of an R^2 :

$\mathrm{var}(y_i|x_i) = \lambda_i = \exp[\beta' x_i] = \sigma_i^2$.

As with the Logit & Probit models, there are several ways of constructing a goodness-of-fit measure —

* The i'th. residual is $(y_i - \tilde{\lambda}_i)$.

Now, $\mathrm{var}(\tilde{y}_i|x_i) = E(\tilde{y}_i|x_i) = \tilde{\lambda}_i$.

So, the i'th. standardized residual is:

$$\left[\frac{y_i - \tilde{\lambda}_i}{\sqrt{\tilde{\lambda}_i}}\right].$$

In the linear regression model:

$$R^2 = 1 - \frac{SSR}{SST}.$$

So, a natural analogue here is:

$$R_p^2 = 1 - \sum_i \left[\frac{y_i - \tilde{\lambda}_i}{\sqrt{\tilde{\lambda}_i}}\right]^2 \Big/ \sum_i \left[\frac{y_i - \bar{y}}{\sqrt{\bar{y}}}\right]^2.$$

(This is also rather like the usual OLS-R^2, in that it compares the fit of the model with what we'd get with just an intercept.

However, unlike usual R^2:

1: It can be negative

2: Its value can fall when we add a regressor to the model.

Other "R^2" measures are also possible, but none of them are perfect.

For example, some packages report the "Likelihood Ratio Index":

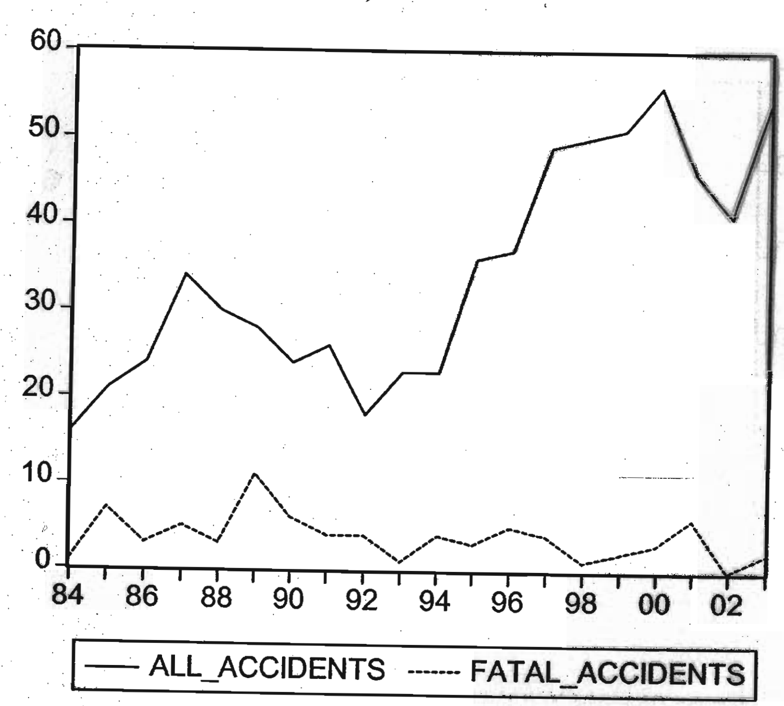
$$R^2_{LRI} = 1 - \left\{\frac{\log L(\hat{\lambda}_i, y_i)}{\log L(\bar{y}, y_i)}\right\}$$

(denominator includes just a constant term)

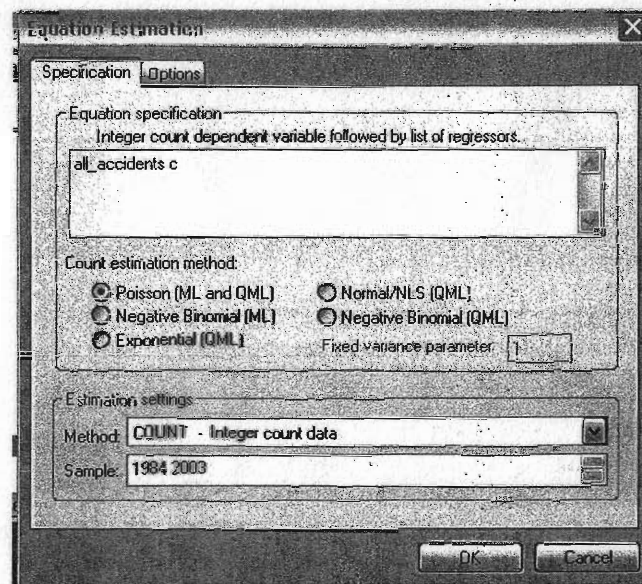
10

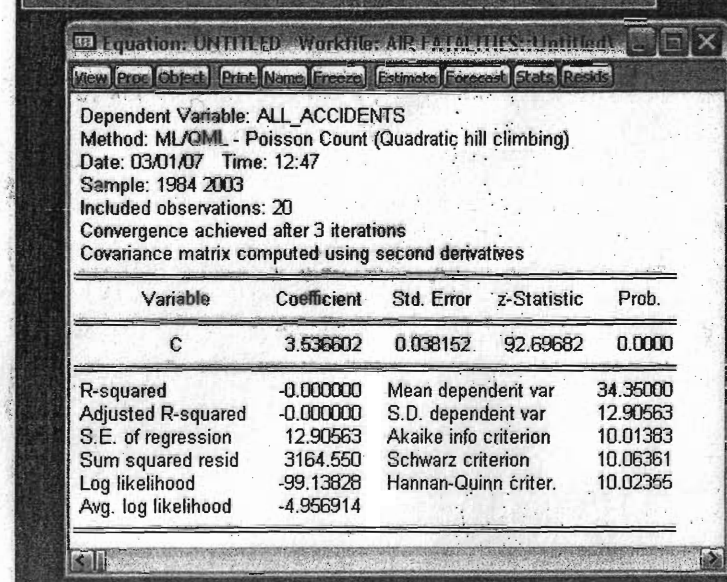
Air Fatalities.xls

USA, 1984-2003.



11





Equation: UNTITLED Workfile: AIR FATALITIES::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: ALL_ACCIDENTS
Method: ML/QML - Poisson Count (Quadratic hill climbing)
Date: 03/01/07 Time: 12:47
Sample: 1984 2003
Included observations: 20
Convergence achieved after 3 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	3.536602	0.038152	92.69682	0.0000

R-squared	-0.000000	Mean dependent var	34.35000
Adjusted R-squared	-0.000000	S.D. dependent var	12.90563
S.E. of regression	12.90563	Akaike info criterion	10.01383
Sum squared resid	3164.550	Schwarz criterion	10.06361
Log likelihood	-99.13828	Hannan-Quinn criter.	10.02355
Avg. log likelihood	-4.956914		

12

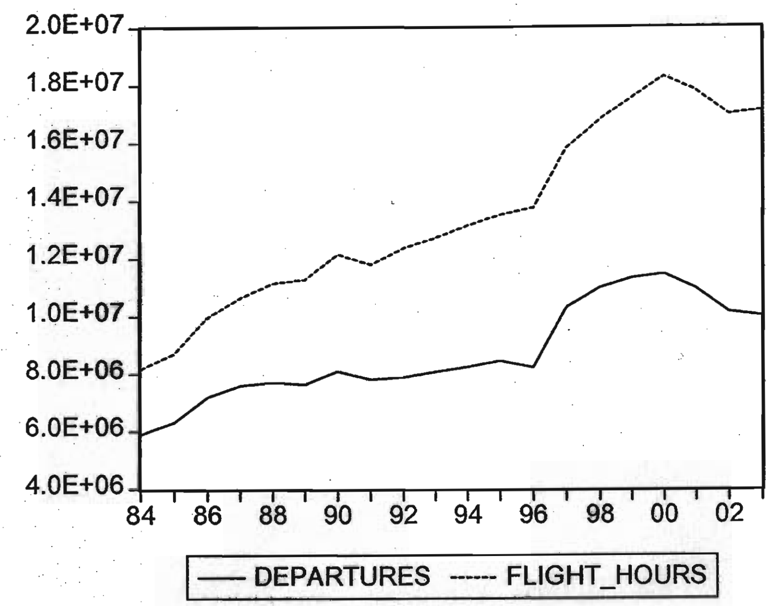
Dependent Variable: ALL_ACCIDENTS
Method: ML/QML - Poisson Count (Quadratic hill climbing)
Date: 06/14/04 Time: 12:27
Sample: 1984 2003
Included observations: 20
Frequencies for dependent variable

Value	Count	Percent	Cumulative Count	Cumulative Percent
16	1	5.00	1	5.00
18	1	5.00	2	10.00
21	1	5.00	3	15.00
23	2	10.00	5	25.00
24	2	10.00	7	35.00
26	1	5.00	8	40.00
28	1	5.00	9	45.00
30	1	5.00	10	50.00
34	1	5.00	11	55.00
36	1	5.00	12	60.00
37	1	5.00	13	65.00
41	1	5.00	14	70.00
46	1	5.00	15	75.00
49	1	5.00	16	80.00
50	1	5.00	17	85.00
51	1	5.00	18	90.00
54	1	5.00	19	95.00
56	1	5.00	20	100.00

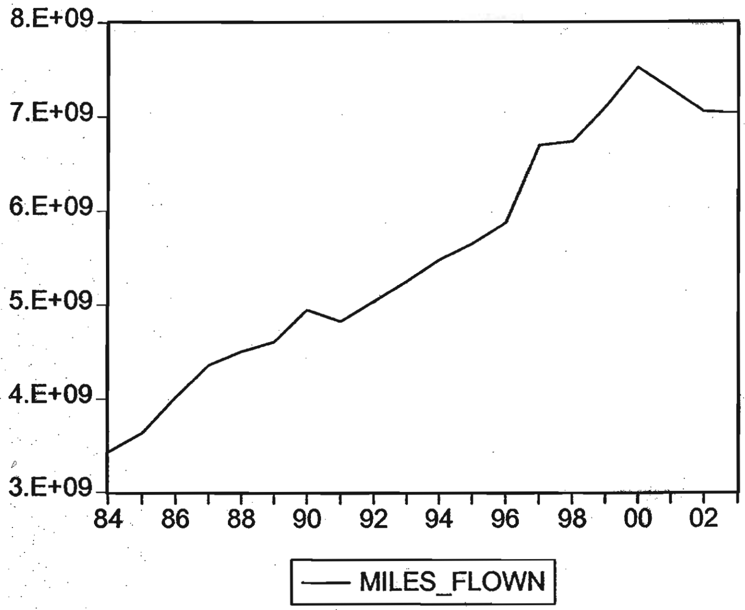
"Counts" for All Accidents.

View / Dependent Variable Frequencies

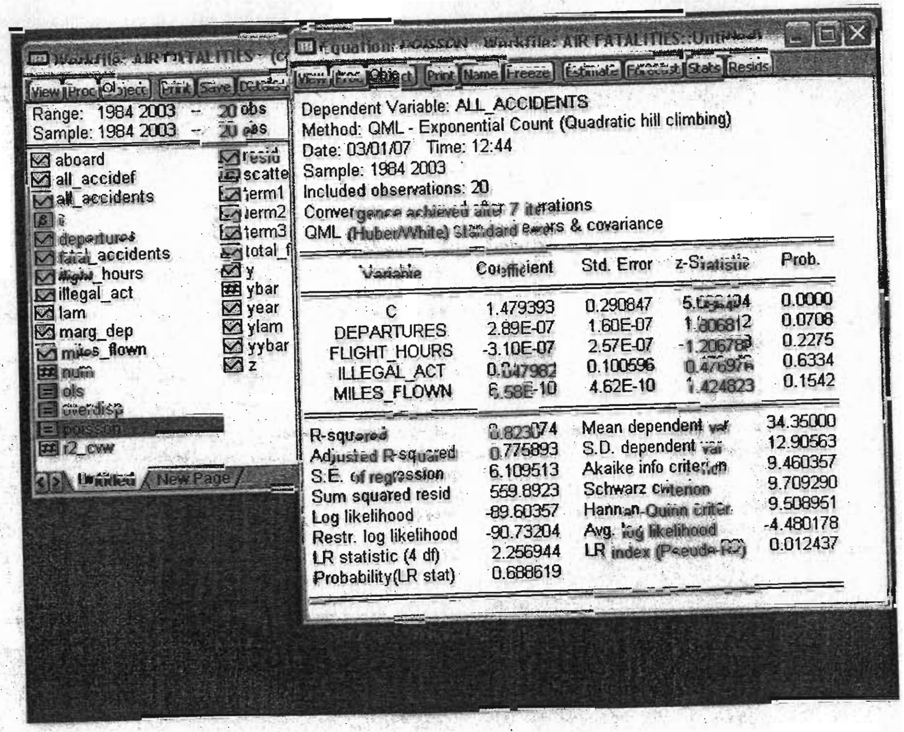
13



14



149



Workfile: AIR FATALITIES

Range: 1984 2003 -- 20 obs
Sample: 1984 2003 -- 20 obs

aboard, all_accidef, all_accidents, c, departures, fatal_accidents, flight_hours, illegal_act, lam, marg_dep, miles_flown, num, ols, overdisp, poisson, r2_cvw, resid, scatte, term1, term2, term3, total_f, y, ybar, year, ylam, yybar, z

Untitled / New Page

Equation: POISSON Workfile: AIR FATALITIES::Untitled

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: ALL_ACCIDENTS
Method: QML - Exponential Count (Quadratic hill climbing)
Date: 03/01/07 Time: 12:44
Sample: 1984 2003
Included observations: 20
Convergence achieved after 7 iterations
QML (Huber/White) standard errors & covariance

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	1.479393	0.290847	[illegible]	0.0000
DEPARTURES	2.89E-07	1.60E-07	1.806812	0.0708
FLIGHT_HOURS	-3.10E-07	2.57E-07	-1.206783	0.2275
ILLEGAL_ACT	0.047982	0.100596	0.476976	0.6334
MILES_FLOWN	6.58E-10	4.62E-10	1.424823	0.1542

R-squared	0.823074	Mean dependent var	34.35000
Adjusted R-squared	0.775893	S.D. dependent var	12.90563
S.E. of regression	6.109513	Akaike info criterion	9.460357
Sum squared resid	559.8923	Schwarz criterion	9.709290
Log likelihood	-89.60357	Hannan-Quinn criter.	9.508951
Restr. log likelihood	-90.73204	Avg. log likelihood	-4.480178
LR statistic (4 df)	2.256944	LR index (Pseudo-R2)	0.012437
Probability(LR stat)	0.688619		

15

Dependent Variable: ALL_ACCIDENTS
Method: Least Squares
Date: 06/14/04 Time: 12:27
Sample: 1984 2003
Included observations: 20
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-26.94373	4.856615	-5.547842	0.0000
DEPARTURES	7.03E-06	5.23E-07	13.43825	0.0000

R-squared	0.826587	Mean dependent var	34.35000
Adjusted R-squared	0.816952	S.D. dependent var	12.90563
S.E. of regression	5.521551	Akaike info criterion	6.349834
Sum squared resid	548.7755	Schwarz criterion	6.449407
Log likelihood	-61.49834	F-statistic	85.79819
Durbin-Watson stat	1.200276	Prob(F-statistic)	0.000000

OLS

Dependent Variable: ALL_ACCIDENTS
Method: ML/QML - Poisson Count (Quadratic hill climbing)
Date: 06/14/04 Time: 12:27
Sample: 1984 2003
Included observations: 20
Convergence achieved after 2 iterations
QML (Huber/White) standard errors & covariance

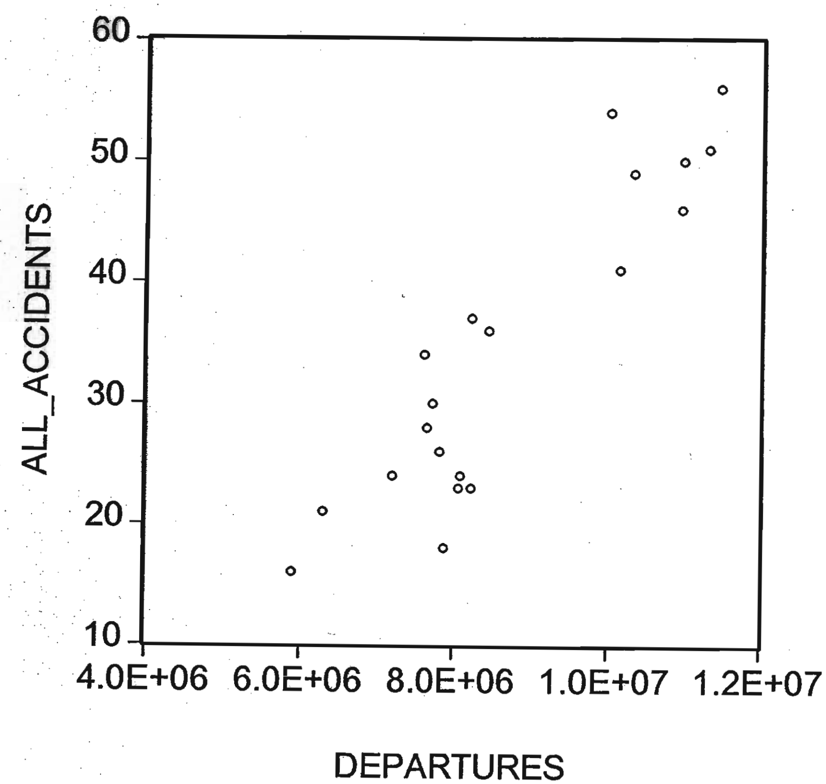
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	1.710210	0.165480	10.33484	0.0000
DEPARTURES	2.03E-07	1.62E-08	12.56127	0.0000

R-squared	0.822298	Mean dependent var	34.35000
Adjusted R-squared	0.812425	S.D. dependent var	12.90563
S.E. of regression	5.589412	Akaike info criterion	6.384373
Sum squared resid	562.3475	Schwarz criterion	6.483946
Log likelihood	-61.84373	Hannan-Quinn criter.	6.403811
Restr. log likelihood	-99.13828	Avg. log likelihood	-3.092186
LR statistic (1 df)	74.58911	LR index (Pseudo-R2)	0.376187
Probability(LR stat)	0.000000		

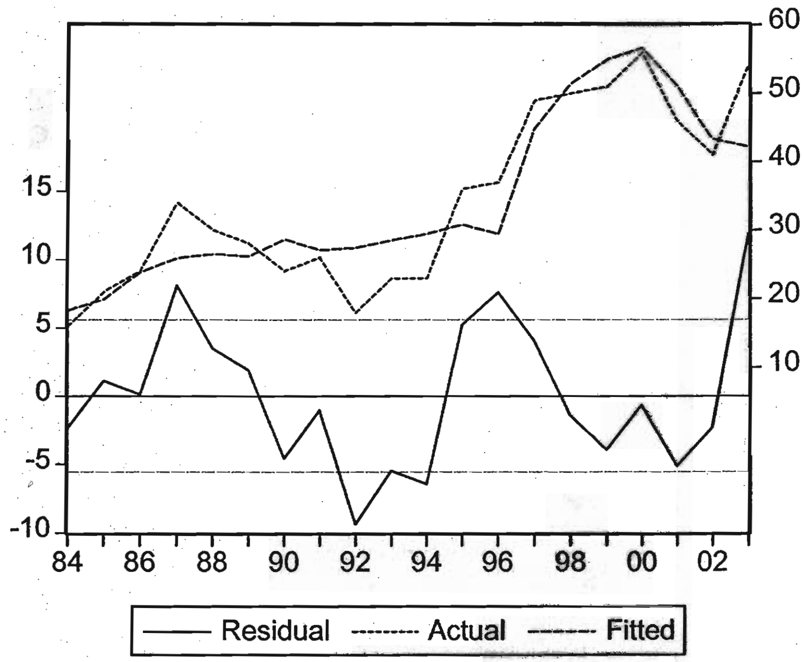
Poisson.

$\leftarrow R^2_{LRI}$

16



17



18.

MARG_DEP

Last updated: 06/15/04 - 12:16
Modified: 1984 2003 // marg_dep=all_accidef*c(2)

1984	3.72E-06
1985	4.04E-06
1986	4.85E-06
1987	5.26E-06
1988	5.38E-06
1989	5.30E-06
1990	5.81E-06
1991	5.49E-06
1992	5.56E-06
1993	5.78E-06
1994	5.98E-06
1995	6.25E-06
1996	5.97E-06
1997	9.13E-06
1998	1.04E-05
1999	1.12E-05
2000	1.15E-05
2001	1.04E-05
2002	8.80E-06
2003	8.56E-06

Marginal Effect for Departures.

$$\frac{\partial E(y_i | \underset{\sim}{x_i})}{\partial D_i} = \tilde{\lambda}_i \tilde{\beta}_2$$

$\boxed{\text{Mean} = 6.97 * 10^{-6}}$

So, if number of departures were reduced by 1,000,000 (10^{+6}) then the decrease in accidents would be approx. 7 in an average year.