Week 10

Methods for Count Data

Dr. Esam Mahdi

Department of Mathematics, Statistics and Physics

Qatar University

Learning Objectives

- Review Poisson distribution.
- Introduce Pearson chi-square goodness of fit test to check the assumptions of the Poisson regression model.
- Study Poisson regression models (Simple and multiple).
- Measure the relative risk associated with an exposure of independent variables.
- Test the hypotheses about the Poisson regression fitted model:
    1. Overall test.
    2. Test for the value of a single factor.
    3. Test for contribution of a group of variables.
- Use statistical software such as R and SPSS to fit Poisson regression model to count data and interpret the output results.

Poisson Distribution

- **The Poisson model** is used when the random variable X represents the number of occurrences (count) of some random event in an interval of time or space, or some volume of matter.
- The **probability density function (*pmf*)** of a Poisson distribution is given by

$$\Pr(X = x) = p(x;\theta) = \frac{\theta^x e^{-\theta}}{x!} \qquad \text{for } x = 0, 1, 2, \ldots$$

- The **mean** and **variance** of the Poisson distribution $P(\theta)$ are

$$\mu = \theta$$

$$\sigma^2 = \theta$$

- We can approximate a Poisson distribution by a normal distribution with mean θ if $\theta \geq 10$

$$z = \frac{x - \theta}{\sqrt{\theta}}$$

Example 10.2

Example 10.2 For the year of 1981, the infant mortality rate (IMR) for the United States was 11.9 deaths per 1000 live births. For the same period, the New England states (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont) had 164,200 live births and 1585 infant deaths. If the national IMR applies, the mean and vaiance of the number of infant deaths in the New England states would be

$$(164.2)(11.9) = 1954$$

From the z score,

$$\begin{aligned} z &= \frac{1585 - 1954}{\sqrt{1954}} \\ &= -8.35 \end{aligned}$$

it is clear that the IMR in the New England states is below the national average.

Testing for Goodness of Fit

- **A goodness-of-fit test** is used when one wishes to decide if an observed distribution of frequencies is incompatible with some hypothesized distribution.
- The Poisson is a special distribution; its mean and its variance are equal.
- The hypotheses are as follows:

H_0: The sampled population is distributed as Poisson.

H_A: The sampled population is not distributed as Poisson.

- Thus, given a sample of count data $\{x_i\}_{i=1}^{n}$, the test statistic is the Pearson chi-square:

$$X^2 = \sum_{i}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i and E_i refer to the ith observed and expected frequencies, respectively.

- k is the number of groups for which observed and expected frequencies are available.
- The null hypothesis will be rejected if the value of the test statistics X^2 is greater than or equal to the tabular value found from the chi-square table at degrees of freedom $k-2$.
- It is recommended that adjacent groups at the bottom of the table be combined to avoid having any expected frequencies less than 1.

Example 10.4

- The purpose of this study was to examine the data for 44 physicians working for an emergency department at a major hospital. The response variable is the number of complaints received during the preceding year.
- The full data is given in Example 10.5 (Table 10.2).

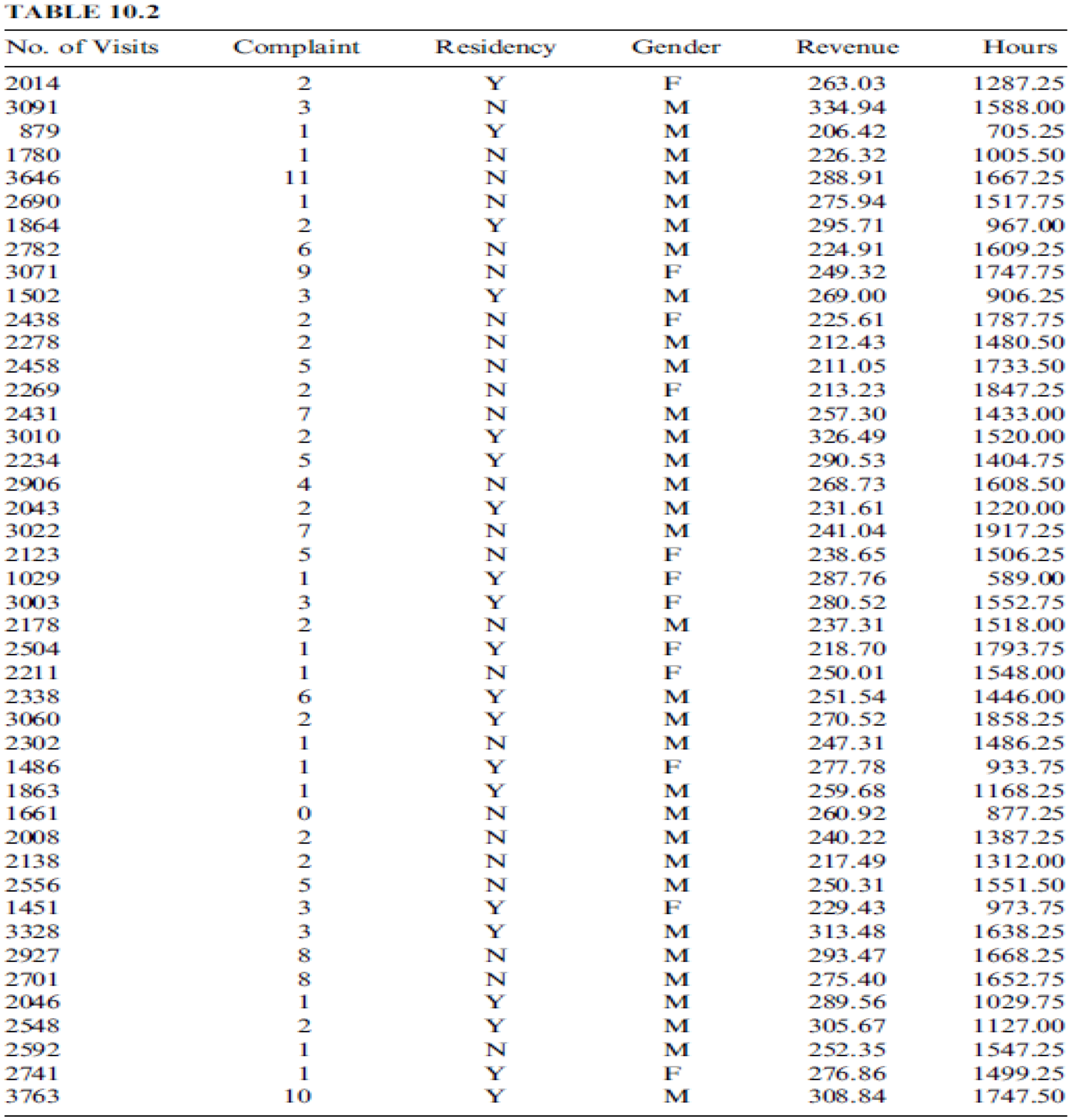


TABLE 10.2

No. of Visits	Complaint	Residency	Gender	Revenue	Hours
2014	2	Y	F	263.03	1287.25
3091	3	N	M	334.94	1588.00
879	1	Y	M	206.42	705.25
1780	1	N	M	226.32	1005.50
3646	11	N	M	288.91	1667.25
2690	1	N	M	275.94	1517.75
1864	2	Y	M	295.71	967.00
2782	6	N	M	224.91	1609.25
3071	9	N	F	249.32	1747.75
1502	3	Y	M	269.00	906.25
2438	2	N	F	225.61	1787.75
2278	2	N	M	212.43	1480.50
2458	5	N	M	211.05	1733.50
2269	2	N	F	213.23	1847.25
2431	7	N	M	257.30	1433.00
3010	2	Y	M	326.49	1520.00
2234	5	Y	M	290.53	1404.75
2906	4	N	M	268.73	1608.50
2043	2	Y	M	231.61	1220.00
3022	7	N	M	241.04	1917.25
2123	5	N	F	238.65	1506.25
1029	1	Y	F	287.76	589.00
3003	3	Y	F	280.52	1552.75
2178	2	N	M	237.31	1518.00
2504	1	Y	F	218.70	1793.75
2211	1	N	F	250.01	1548.00
2338	6	Y	M	251.54	1446.00
3060	2	Y	M	270.52	1858.25
2302	1	N	M	247.31	1486.25
1486	1	Y	F	277.78	933.75
1863	1	Y	M	259.68	1168.25
1661	0	N	M	260.92	877.25
2008	2	N	M	240.22	1387.25
2138	2	N	M	217.49	1312.00
2556	5	N	M	250.31	1551.50
1451	3	Y	F	229.43	973.75
3328	3	Y	M	313.48	1638.25
2927	8	N	M	293.47	1668.25
2701	8	N	M	275.40	1652.75
2046	1	Y	M	289.56	1029.75
2548	2	Y	M	305.67	1127.00
2592	1	N	M	252.35	1547.25
2741	1	Y	F	276.86	1499.25
3763	10	Y	M	308.84	1747.50

Example 10.4 (Cont.)

- For the purpose of testing the goodness of fit, the data are summarized in Table 10.1.
- The expected relative frequency $E(x) = nP(X = x)$, where n is the sample size and $nP(X = x)$ is the relative frequency obtained by evaluating the Poisson probability for the value of $X = \mathrm{x}$.

$$\Pr(X = x) = \frac{\theta^x e^{-\theta}}{x!} \qquad \text{for } x = 0, 1, 2, \ldots$$

Since $\hat{\theta} = \frac{\sum x_i}{44} = \frac{0\times1+1\times12+\cdots+11\times1}{44} = 3.44$, we have for example,

$$\Pr(X = 2) = \frac{3.34^2 e^{-3.34}}{2!} = 0.198 \text{ and } E_2 = (44)(0.198) = 8.71$$

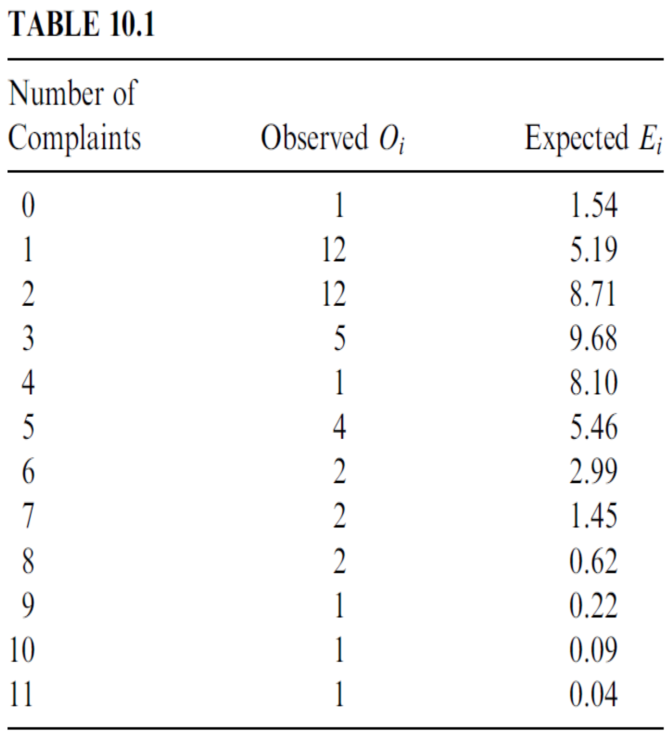


TABLE 10.1

Number of Complaints	Observed O_i	Expected E_i
0	1	1.54
1	12	5.19
2	12	8.71
3	5	9.68
4	1	8.10
5	4	5.46
6	2	2.99
7	2	1.45
8	2	0.62
9	1	0.22
10	1	0.09
11	1	0.04

- To avoid having any expected frequencies less than 1, we combine the last five groups together, resulting in eight groups available for testing goodness of fit with

$$O_8 = 2 + 2 + 1 + 1 + 1 = 7$$

$$E_8 = 1.45 + 0.62 + 0.22 + 0.09 + 0.04 = 2.42$$

The result is $X^2 = \frac{(1 - 1.54)^2}{1.19} + \frac{(12 - 5.19)^2}{5.19} + \cdots + \frac{(7 - 2.42)^2}{2.42} = 28.24$

with $8 - 2 = 6$ degrees of freedom, indicating a significant deviation from the Poisson distribution ($p < 0.005$).

Poisson Regression Model

- The response (dependent) variable Y is a count variable (e.g., the number of defective teeth per person, the number of patients with HIV, etc.) follows a Poisson distribution.
- The Poisson regression model is a function of other variables $X_1, X_2, \ldots, X_k$ in addition to the size of the observation unit from which one obtained the count of interest.
  - For example, if Y is the number of virus in a solution, the size is the volume of the solution.
  - If Y is the number of defective teeth, the size is the total number of teeth for that same person.
- The other variables $X_1, X_2, \ldots, X_k$ are called predictors, or, explanatory variables, or independent variables, or covariates.

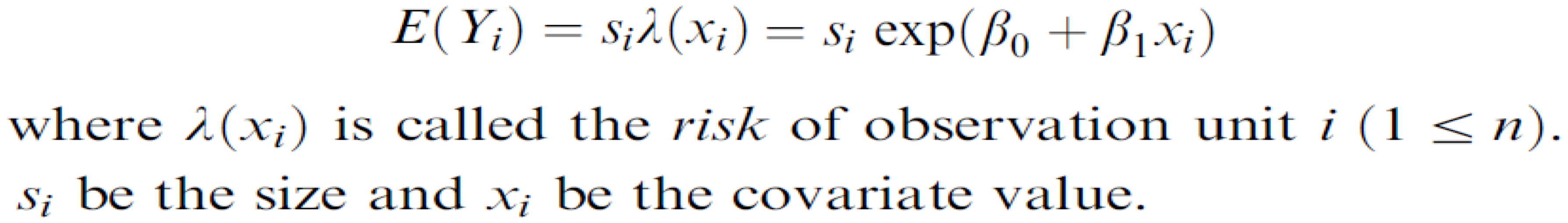
Example 10.5

- The purpose of this study was to examine the data for 44 physicians working for an emergency at a major hospital so as to determine which of four variables are related to the number of complaints received during the preceding year.
- Table 10.2 presents the complete data set.
- Data available consist of the number of visits (which serves as the size for the observation unit, the physician) and four covariates. In addition to the number of complaints is the dependent variable.
- The number of complaints is the dependent variable.
- For each of the 44 physicians there are two continuous independent variables, the revenue (dollars per hour) and the workload at the emergency service (hours) and two binary variables, gender (female/male) and residency training in emergency services (no/yes).

Simple Regression Analysis (Poisson Model)

Assumptions:

- The dependent variable Y is assumed to follow a Poisson distribution.
- The Poisson regression model assumes that the relationship between the mean of Y and the covariate X is described by



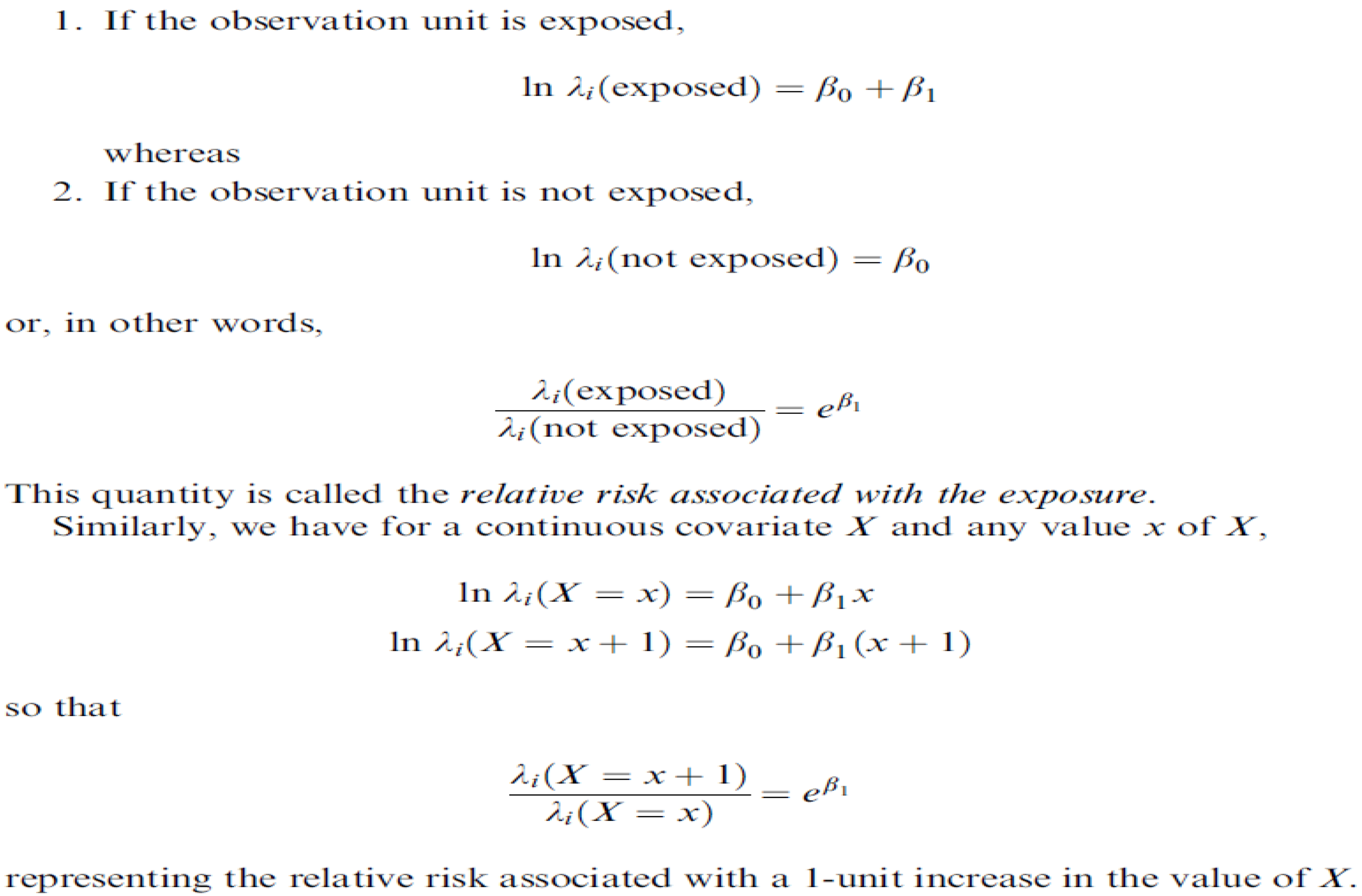
$$E(Y_i) = s_i \lambda(x_i) = s_i \exp(\beta_0 + \beta_1 x_i)$$

where $\lambda(x_i)$ is called the *risk* of observation unit i $(1 \leq n)$.
s_i be the size and x_i be the covariate value.

- Under the assumption that $Y_i, i = 1, \dots, n$ is Poisson, the estimators of the regression coefficients β_0 and $\beta_1$can be derived from the maximum likelihood procedure.

Measure of Association

- Consider the case of a binary covariate X: say, representing an exposure (1 = exposed, 0 = not exposed). We have:



1. If the observation unit is exposed,

$$\ln \lambda_i(\text{exposed}) = \beta_0 + \beta_1$$

whereas

2. If the observation unit is not exposed,

$$\ln \lambda_i(\text{not exposed}) = \beta_0$$

or, in other words,

$$\frac{\lambda_i(\text{exposed})}{\lambda_i(\text{not exposed})} = e^{\beta_1}$$

This quantity is called the *relative risk associated with the exposure*.

Similarly, we have for a continuous covariate X and any value x of X,

$$\ln \lambda_i(X = x) = \beta_0 + \beta_1 x$$
$$\ln \lambda_i(X = x + 1) = \beta_0 + \beta_1 (x + 1)$$

so that

$$\frac{\lambda_i(X = x + 1)}{\lambda_i(X = x)} = e^{\beta_1}$$

representing the relative risk associated with a 1-unit increase in the value of X.

Example 10.6

- The emergency service data in Example 10.5 (Table 10.2).
- Investigate the relationship between the number of complaints (adjusted for number of visits) and residency training (It may be perceived that by having training in the specialty a physician would perform better and therefore would be less likely to provoke complaints.)

Solution:

- The simple Poisson regression analysis yields the results shown in Table 10.3.

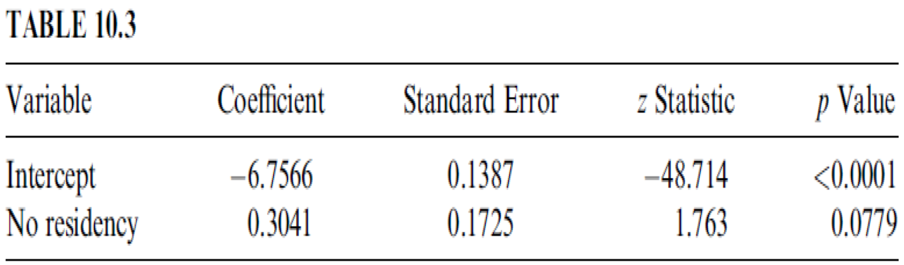


TABLE 10.3

Variable	Coefficient	Standard Error	z Statistic	p Value
Intercept	−6.7566	0.1387	−48.714	<0.0001
No residency	0.3041	0.1725	1.763	0.0779

- The result indicates that the relationship between the number of complaints and no residency training in emergency service is marginally significant ($p - value = 0.0779$);
- the relative risk associated with no residency training is

$$\exp(\beta_1) = \exp(0.3041) = 1.36$$

- Those without previous training is 36% more likely to receive the same number of complaints as those who were trained in the specialty.

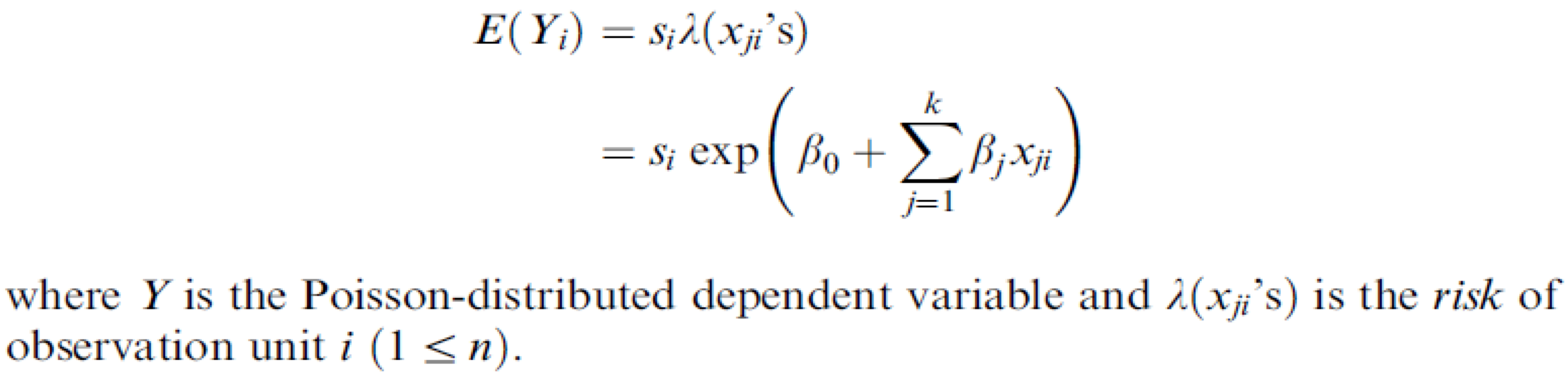
Multiple Regression Analysis (Poisson Model)

Assumptions:

- The dependent variable Y is assumed to follow a Poisson distribution.
- The multiple Poisson regression analysis, involves a linear combination of the explanatory or independent variables;
- The variables must be quantitative with particular numerical values for each observation unit.
- A covariate or independent variable may be dichotomous, polytomous, or continuous; categorical factors will be represented by dummy variables (0 or 1).
- In many cases, data transformations of continuous measurements (e.g., taking the logarithm) may be desirable so as to satisfy the linearity assumption.

Poisson Regression Model with Several Covariates

- Suppose that we want to consider k covariates, $X_1, X_2, \dots, X_k$, simultaneously.
- The multiple Poisson regression model can be expressed as



$$\begin{aligned} E(Y_i) &= s_i \lambda(x_{ji}\text{'s}) \\ &= s_i \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}\right) \end{aligned}$$

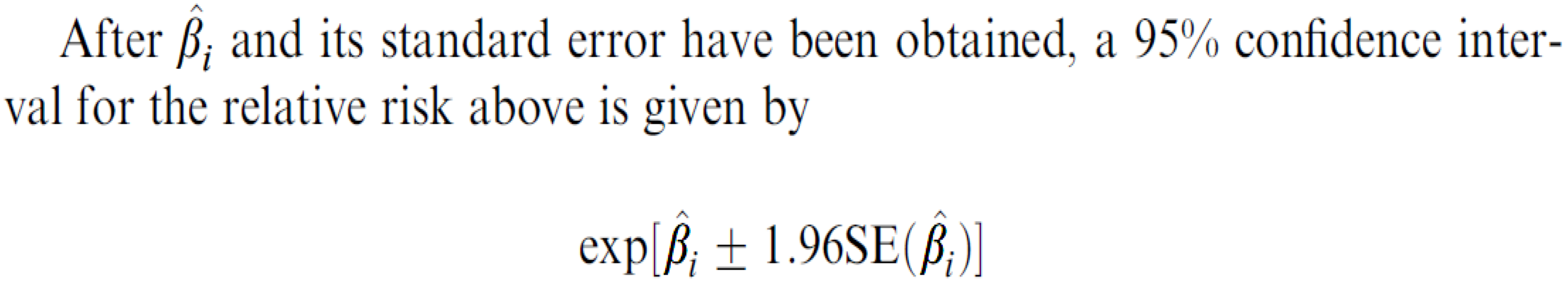
where Y is the Poisson-distributed dependent variable and $\lambda(x_{ji}\text{'s})$ is the *risk* of observation unit i $(1 \leq n)$.

- In many cases, we may include the interactions between independent variables in the model (for example, when $X_1$and X_2 interacted, the Poisson model can be expressed as $\mathrm{E}(Y_i) = s_i(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2)$ where $\beta_3 x_1 x_2$ is the interaction term.)
- Under the assumption that Y_i is Poisson, the estimators of $\beta_0, \beta_1, \dots, \beta_k$ can be obtained by the maximum likelihood procedure.

The relative risk associated with an exposure

- Similar to the simple regression case $\exp(\beta_\mathrm{i})$ represents:
  1. The relative risk associated with an exposure if X_i is binary (exposed $X_i = 1$ versus unexposed $X_i = 0$), or
  2. The relative risk due to a 1-unit increase if X_i is continuous ($X_i = x + 1$ versus $X_i = x$).

Confidence interval for the relative risk



After $\hat{\beta}_i$ and its standard error have been obtained, a 95% confidence interval for the relative risk above is given by

$$\exp[\hat{\beta}_i \pm 1.96\mathrm{SE}(\hat{\beta}_i)]$$

Testing Hypotheses in Multiple Poisson Regression

There are three types of testing hypotheses:

1. **Overall test**. Taken collectively, does the entire set of explanatory or independent variables contribute significantly to the prediction of response?
2. **Test for the value of a single factor**. Does the addition of one particular variable of interest add significantly to the prediction of response over and above that achieved by other independent variables?
3. **Test for contribution of a group of variables**. Does the addition of a group of variables add significantly to the prediction of response over and above that achieved by other independent variables?

Overall Regression Test

- The null hypothesis for this test may stated as: ‘‘All k independent variables considered together do not explain the variation in the response any more than the size alone.’’

$$H_0 \colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

- This can be tested using the likelihood ratio chi-square test at k degrees of freedom:

$$\chi^2 = 2(\ln L_k - \ln L_0)$$

- where $\ln L_k$ is the log likelihood value for the model containing all k covariates (full model) and $\ln L_0$ is the log likelihood value for the model containing only the intercept (null model).

Example 10.7

- Refer to the data set on emergency service of Example 10.5 (Table 10.2) with four covariates: gender, residency, revenue, and workload (hours):
- The null hypothesis: $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
  - For all four covariates included, $\ln L_4 = 47.783$, whereas
  - With no covariates included, $\ln L_0 =$ 43.324.
- $\chi^2 = 8.918$ with 4 degrees of freedom (p-value = 0.0636) indicating that at least one covariate must be moderately related significantly to the number of complaints.

Test for a Single Variable

- The null hypothesis for this test may be stated as: ‘‘Factor X_i does not have any value added to the prediction of the response given that other factors are already included in the model.’’

$$H_0\colon \beta_i = 0$$

- To test such a null hypothesis, one can use

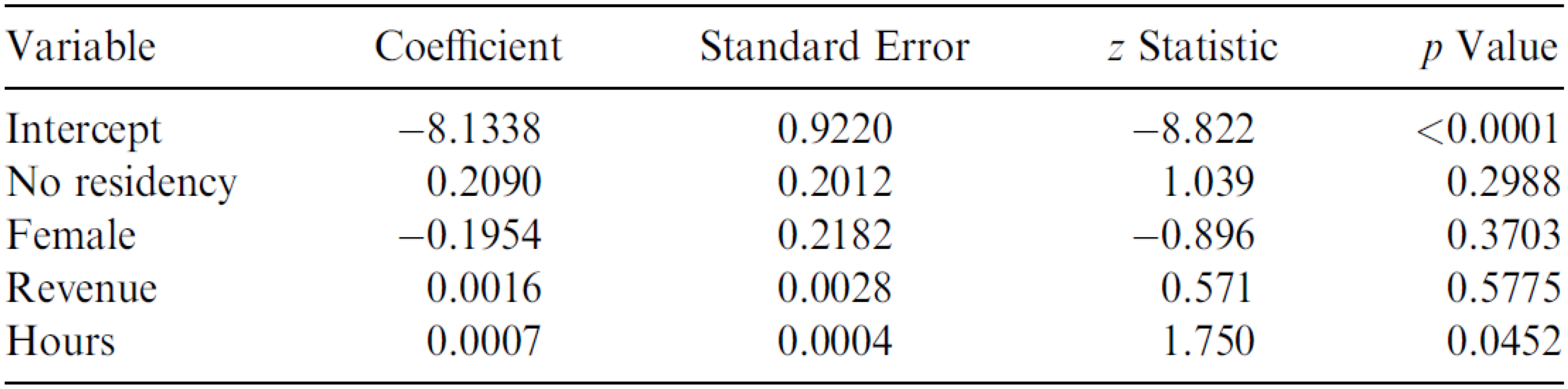
$$z_i = \frac{\hat{\beta}_i}{\mathrm{SE}(\hat{\beta}_i)}$$

- where $\hat{\beta}_i$ is the corresponding estimated regression coefficient and $SE(\hat{\beta}_i)$ is the estimate of the standard error of $\hat{\beta}_i$,
- We refer the value of the z score; to percentiles of the standard normal distribution.
  - for example, we reject the null hypothesis if the absolute value of z is greater than or equal to 1.96 for a two-sided test at the 5% level.

Example 10.8

- Refer to the data set on emergency service of Example 10.5 (Table 10.2) with all four covariates. We have the results shown in Table 10.4.

TABLE 10.4



Variable	Coefficient	Standard Error	*z* Statistic	*p* Value
Intercept	–8.1338	0.9220	–8.822	<0.0001
No residency	0.2090	0.2012	1.039	0.2988
Female	–0.1954	0.2182	–0.896	0.3703
Revenue	0.0016	0.0028	0.571	0.5775
Hours	0.0007	0.0004	1.750	0.0452

- Only the effect of workload (hours) is significant at the 5% level.

Example 10.10

- The dependent variable is the number of cases of skin cancer.
- Data were obtained from two metropolitan areas:
  - Minneapolis-St. Paul from south.
  - Dallas-Ft. Worth from north.
- The population of each area is divided into eight age groups.
- Data are shown in Table 10.5.
- We need to compare the incidences of nonmelanoma skin cancer among women from two the major metropolitan areas.

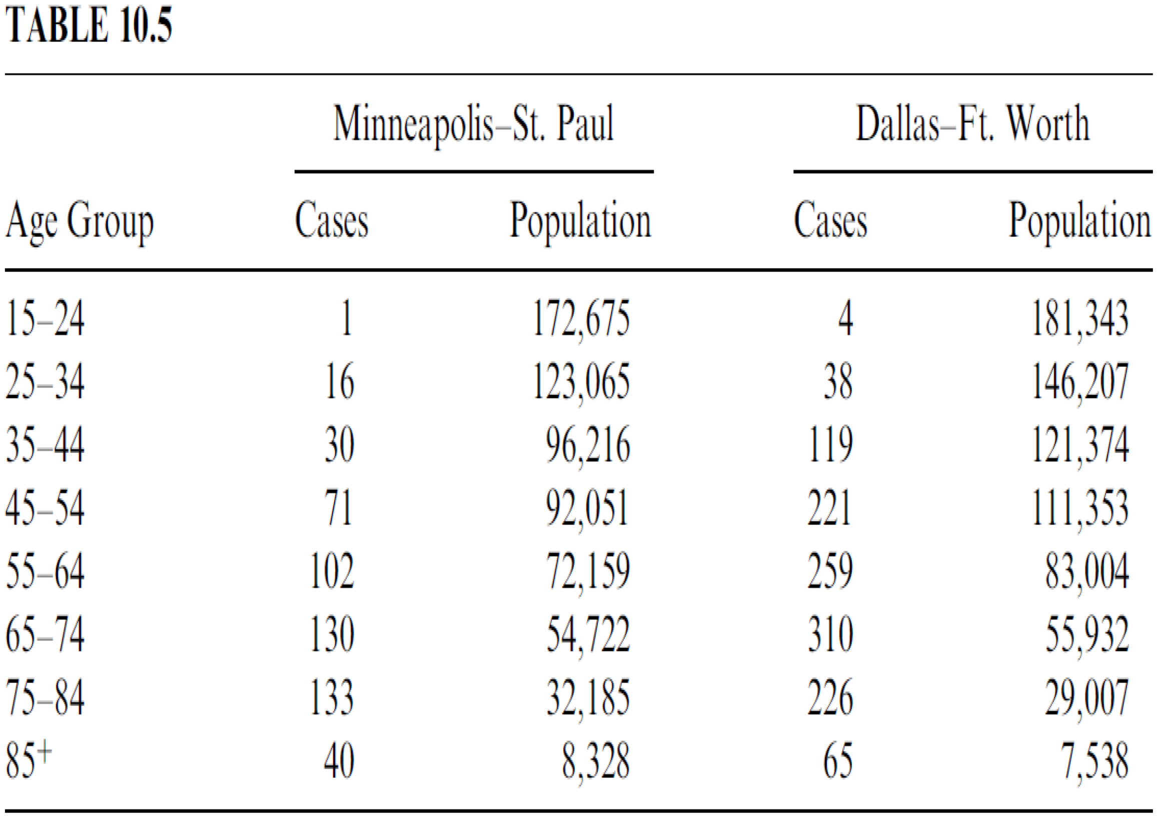


TABLE 10.5

	Minneapolis–St. Paul		Dallas–Ft. Worth	
Age Group	Cases	Population	Cases	Population
15–24	1	172,675	4	181,343
25–34	16	123,065	38	146,207
35–44	30	96,216	119	121,374
45–54	71	92,051	221	111,353
55–64	102	72,159	259	83,004
65–74	130	54,722	310	55,932
75–84	133	32,185	226	29,007
85+	40	8,328	65	7,538

Example 10.10 (Cont.)

- This problem involves two covariates: age and location; both are categorical.
  - We use seven dummy variables to represent the eight age groups (with 85+ being the baseline),
  - and one for location (with Minneapolis-St. Paul as the baseline.
  - The results in Table 10.6.
- The results indicate an upward trend of skin cancer incidence with age, and with Minneapolis-St. Paul as the baseline.
- The relative risk associated with Dallas-Ft. Worth is

  relative risk = exp(0.8043) = 2.235

an increase of more than twofold for this southern metropolitan area.

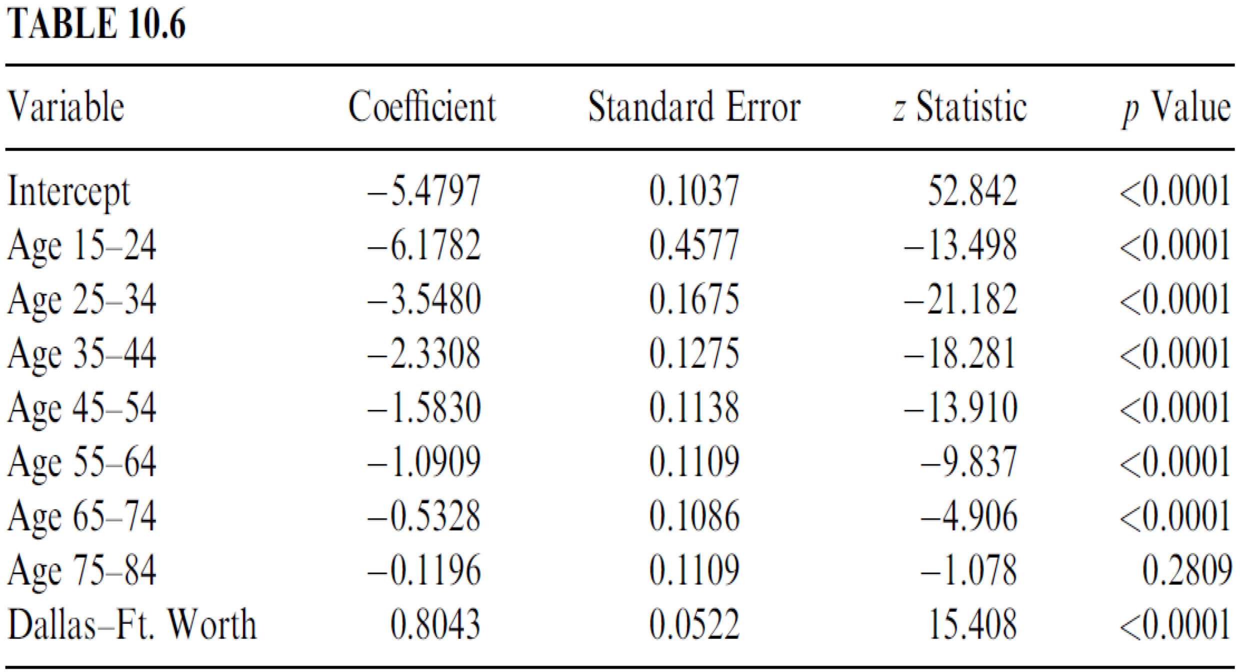


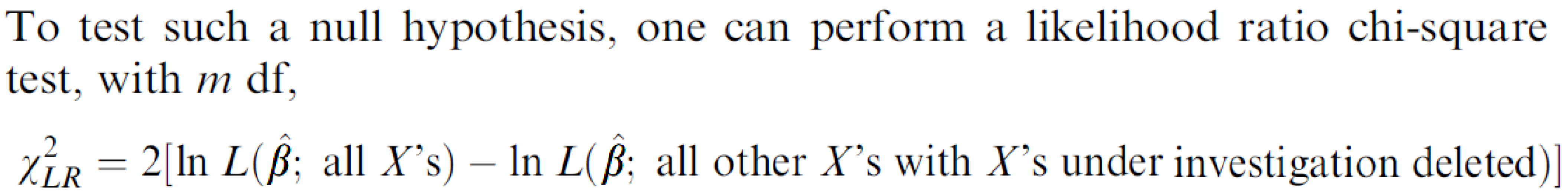
TABLE 10.6

Variable	Coefficient	Standard Error	*z* Statistic	*p* Value
Intercept	−5.4797	0.1037	52.842	<0.0001
Age 15–24	−6.1782	0.4577	−13.498	<0.0001
Age 25–34	−3.5480	0.1675	−21.182	<0.0001
Age 35–44	−2.3308	0.1275	−18.281	<0.0001
Age 45–54	−1.5830	0.1138	−13.910	<0.0001
Age 55–64	−1.0909	0.1109	−9.837	<0.0001
Age 65–74	−0.5328	0.1086	−4.906	<0.0001
Age 75–84	−0.1196	0.1109	−1.078	0.2809
Dallas–Ft. Worth	0.8043	0.0522	15.408	<0.0001

Contribution of a Group of Variables

- This testing procedure addresses the more general problem of assessing the additional contribution of two or more factors to the prediction of the response over and above that made by other variables already in the regression model.
- The null hypothesis is of the form

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_m = 0$$



To test such a null hypothesis, one can perform a likelihood ratio chi-square test, with m df,

$$\chi^2_{LR} = 2[\ln L(\hat{\beta};\ \text{all } X\text{'s}) - \ln L(\hat{\beta};\ \text{all other } X\text{'s with } X\text{'s under investigation deleted})]$$

- This multiple contribution procedure is often used to test whether a similar group of variables, such as demographic characteristics, is important for the prediction of the response; these variables have some trait in common. Another application would be a collection of powers and/or product terms (referred to as interaction variables).

Example 10.11

- Refer to the data set on skin cancer of Example 10.10 (Table 10.5) with all eight covariates, and we consider collectively the 7 dummy variables representing the age.
- The basic idea is to see if there are any differences without drawing seven separate conclusion comparing each age group versus the baseline.
- For all 8 covariates included, $\ln L = 7201.864$, whereas
- With the 7 age variables were deleted $(\beta_1 = \beta_2 = \cdots = \beta_7 = 0)$, $\ln L = 5921.076$.
- $\chi^2 = 2[7201.864 - 5921.076] = 2561.576$ with 7 degrees of freedom $(p - value \ < 0.0001)$ indicating that at least one covariate must be moderately related significantly to the number of complaints.
- In other words, the difference between the age group is highly significant; in fact, it is more so than the difference between the cities.