Poisson Regression
or
Regression of Counts (& Rates)

Carolyn J. Anderson

Department of Educational Psychology

University of Illinois at Urbana-Champaign

Outline

- GLMs for count data.
  - Poisson regression for counts.
  - Poisson regression for rates.
- Inference and model checking.
  - Wald, Likelihood ratio, & Score test.
  - Checking Poisson regression.
  - Residuals.
  - Confidence intervals for fitted values (means).
  - Overdispersion.
- Fitting GLMS (a little technical).
  - Newton-Raphson algorithm/Fisher scoring.
  - Statistic inference & the Likelihood function.
  - "Deviance".
- Summary

GLMs for count data

Situation: response/outcome variable Y is a count.

Generalized linear models for counts have as it's random component Poisson Distribution.

Examples:

- Number of cargo ships damaged by waves (classic example given by McCullagh & Nelder, 1989).
- Number of deaths due to AIDs in Australia per quarter (3 month periods) from January 1983 – June 1986.
- Number of violent incidents exhibited over a 6 month period by patients who had been treated in the ER of a psychiatric hospital (Gardner, Mulvey, & Shaw, 1995).
- Daily homicide counts in California (Grogger, 1990).
- Foundings of day care centers in Toronto (Baum & Oliver, 1992).
- Political party switching among members of the US House of Representatives (King, 1988).

More Examples...

- Number of presidential appointments to the Supreme Court (King, 1987).
- Number of children in a classroom that a child lists as being their friend (unlimited nomination procedure, sociometric data).
- Number of hard disk failures at uiuc during a year.
- Number of deaths due to SARs (Yu, Chan & Fung, 2006).
- Number of arrests resulting from 911 calls.
- Number of orders of protection issued.

In some of these examples, we should consider "exposure" to the event. i.e., "t".

e.g., hard disk failures: In this case, "exposure" could be the number of hours of operation. Rather than model the number of failures (i.e., counts), we would want to measure and model the failure "rate"

$$Y/t = \text{rate}$$

Poisson regression for counts

Response Variable is a count

Explanatory Variable(s):

- If they are categorical (i.e., you have a contingency table with counts in the cells), convention is to call them "Loglinear models".
- If they are numerical/continuous, convention is to call them "Poisson Regression"

First, $Y =$ count and then Y/t rate data.

Comonents of GLM for Counts

- Random component: Poisson distribution and model the expected value of Y, denoted by $E(Y) = \mu$.
- Systematic component: For now, just 1 explanatory variable x (later, we'll go over an example with more than 1).
- Link: We could use
  - Identity link, which gives us $\mu = \alpha + \beta x$

    Problem: a linear model can yield $\mu < 0$, while the possible values for $\mu \geq 0$.
  - Log link (much more common) $\log(\mu)$. which is the "natural parameter" of Poisson distribution, and the log link is the "canonical link" for GLMs with Poisson distribution.

The Poisson regression model for counts (with a log link) is

$$\log(\mu) = \alpha + \beta x$$

This is often referred to as "Poisson loglinear model".

The Poisson loglinear model

$$\log(\mu) = \alpha + \beta x$$

Since the log of the expected value of Y is a linear function of explanatory variable(s), and the expected value of Y is a multiplicative function of x:

$$\begin{aligned} \mu &= \exp(\alpha + \beta x) \\ &= e^{\alpha} e^{\beta x} \end{aligned}$$

What does this mean for μ?

How do we interpret β?

Interpretation of β

$$\log(\mu) = \alpha + \beta x$$

Consider 2 values of x (x_1 & x_2) such that the difference between them equals 1. For example, $x_1 = 10$ and $x_2 = 11$:

$$x_2 = x_1 + 1$$

The expected value of μ when $x = 10$ is

$$\mu_1 = e^{\alpha} e^{\beta x_1} = e^{\alpha} e^{\beta(10)}$$

The expected value of μ when $x = x_2 = 11$ is

$$\begin{aligned} \mu_2 &= e^{\alpha} e^{\beta x_2} \\ &= e^{\alpha} e^{\beta(x_1+1)} \\ &= e^{\alpha} e^{\beta x_1} e^{\beta} \\ &= e^{\alpha} e^{\beta(10)} e^{\beta} \end{aligned}$$

A change in x has a multiplicative effect on the mean of Y.

Interpretation of β (continued)

When we look at a 1 unit increase in the explanatory variable (i.e., $x_2 - x_1 = 1$), we have

$$\mu_1 = e^{\alpha} e^{\beta x_1} \qquad \text{and} \qquad \mu_2 = e^{\alpha} e^{\beta x_1} e^{\beta}$$

- If $\beta = 0$, then $e^0 = 1$ and
  - $\mu_1 = e^{\alpha}$.
  - $\mu_2 = e^{\alpha}$.
  - $\mu = E(Y)$ is not related to x.
- If $\beta > 0$, then $e^{\beta} > 1$ and
  - $\mu_1 = e^{\alpha} e^{\beta x_1}$
  - $\mu_2 = e^{\alpha} e^{\beta x_2} = e^{\alpha} e^{\beta x_1} e^{\beta} = \mu_1 e^{\beta}$
  - μ_2 is e^{β} times larger than μ_1.
- If $\beta < 0$, then $0 \leq e^{\beta} < 1$
  - $\mu_1 = e^{\alpha} e^{\beta x_1}$.
  - $\mu_2 = e^{\alpha} e^{\beta x_2} = e^{\alpha} e^{\beta x_1} e^{\beta} = \mu_1 e^{\beta}$.
  - μ_2 is e^{β} times smaller than μ_1.

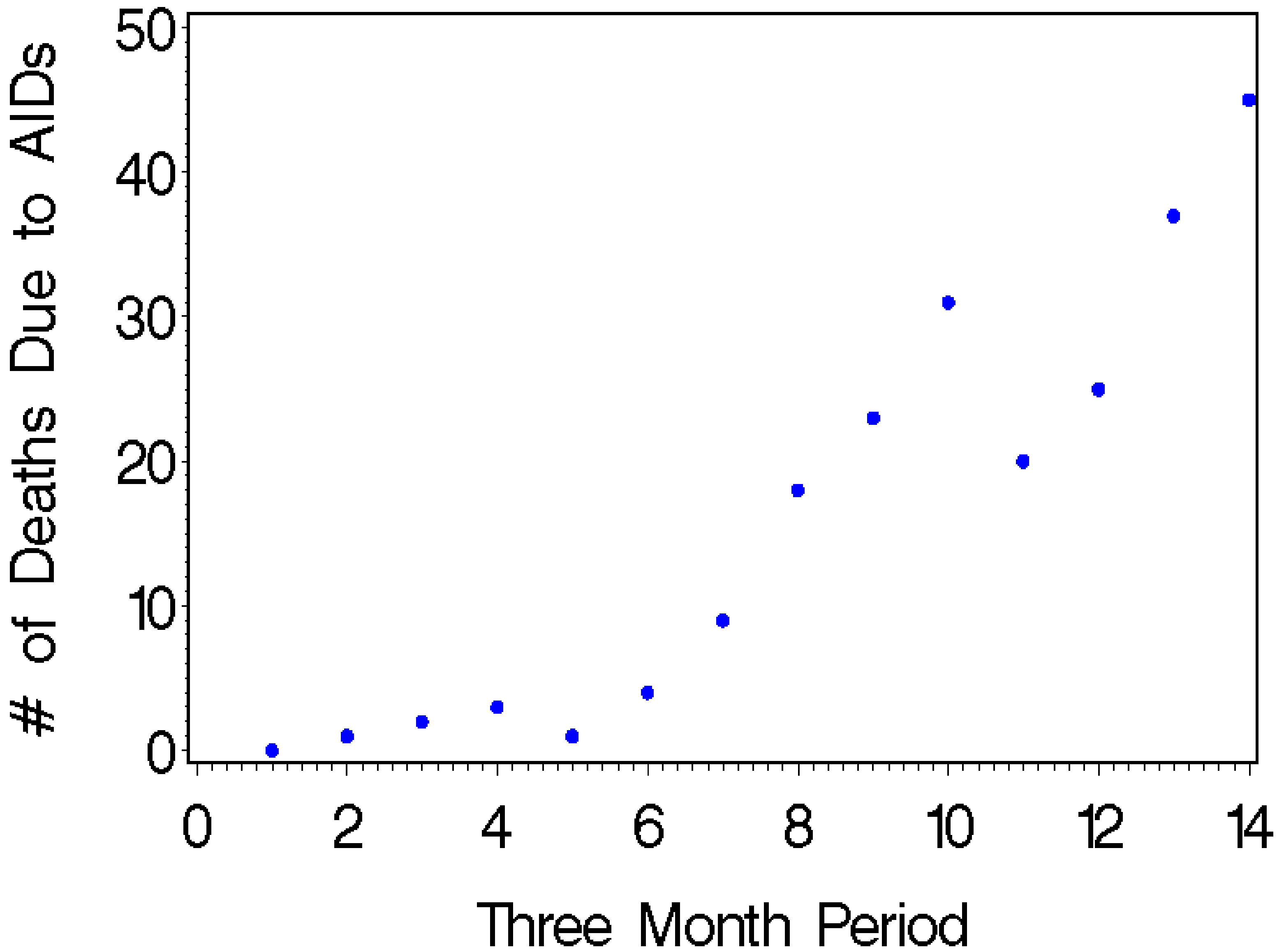
Example: Number of Deaths Due to AIDs

Whyte, et al 1987 (Dobson, 1990) reported the number of deaths due to AIDS in Australia per 3 month period from January 1983 – June 1986.

$y_i =$ number of deaths
$x_i =$ time point (quarter)

x_i	y_i	x_i	y_i
1	0	8	18
2	1	9	23
3	2	10	31
4	3	11	20
5	1	12	25
6	4	13	37
7	9	14	45

Data: Number of Deaths Due to AIDs $\times$ Month



A Linear Model for AIDs Data

Let's try a linear model:

$$\mu_i = \alpha + \beta x_i$$

The estimated parameters from GLM with a Poisson distribution and the identity link:

$$\hat{\mu}_i = -6.7355 + 2.4287 x_i$$

In SAS OUTPUT, there's *strange* things such as

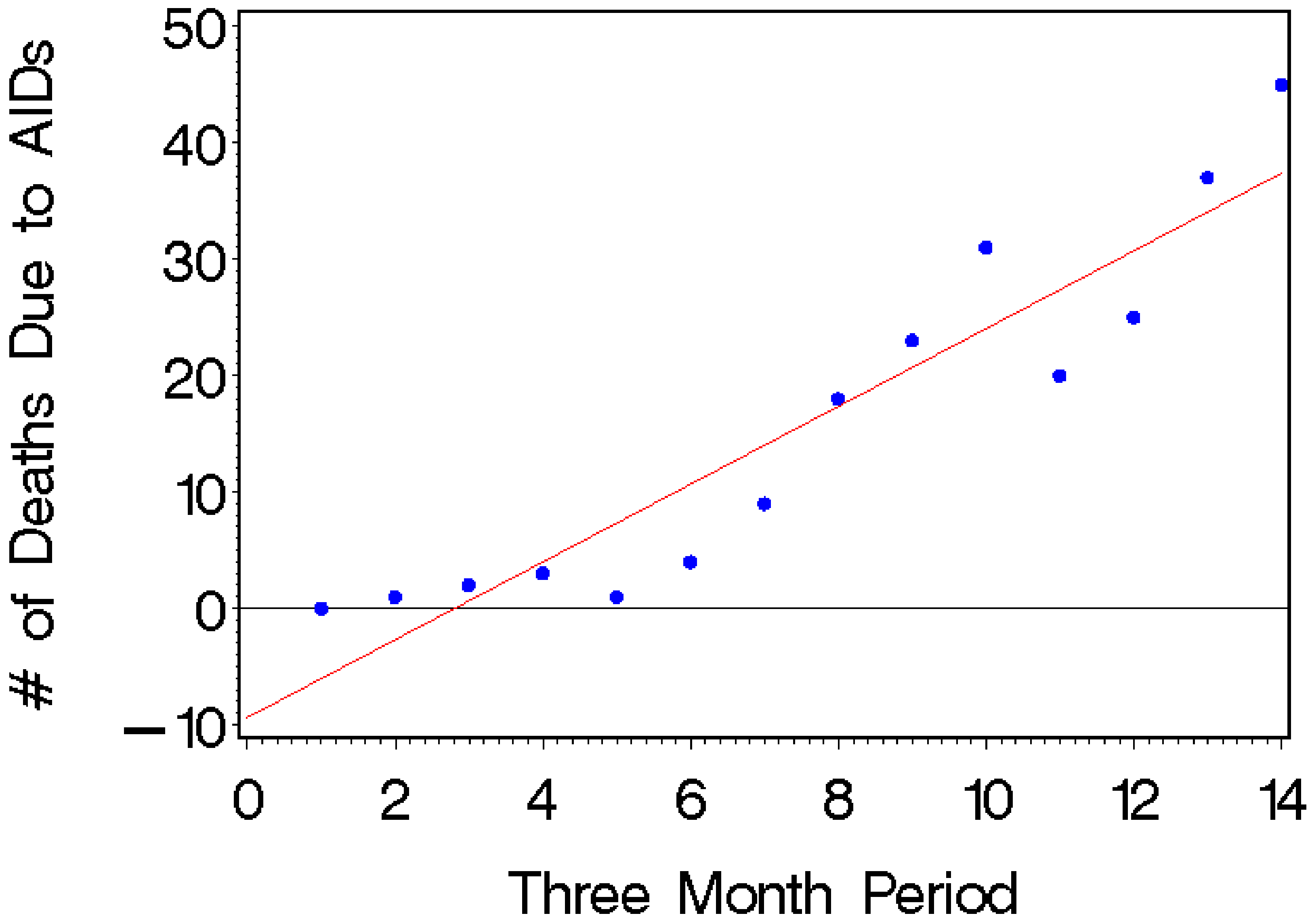
- Standard errors for estimated parameters equal to 0.
- Some 0's in the OBSTATS.

From SAS LOG file...

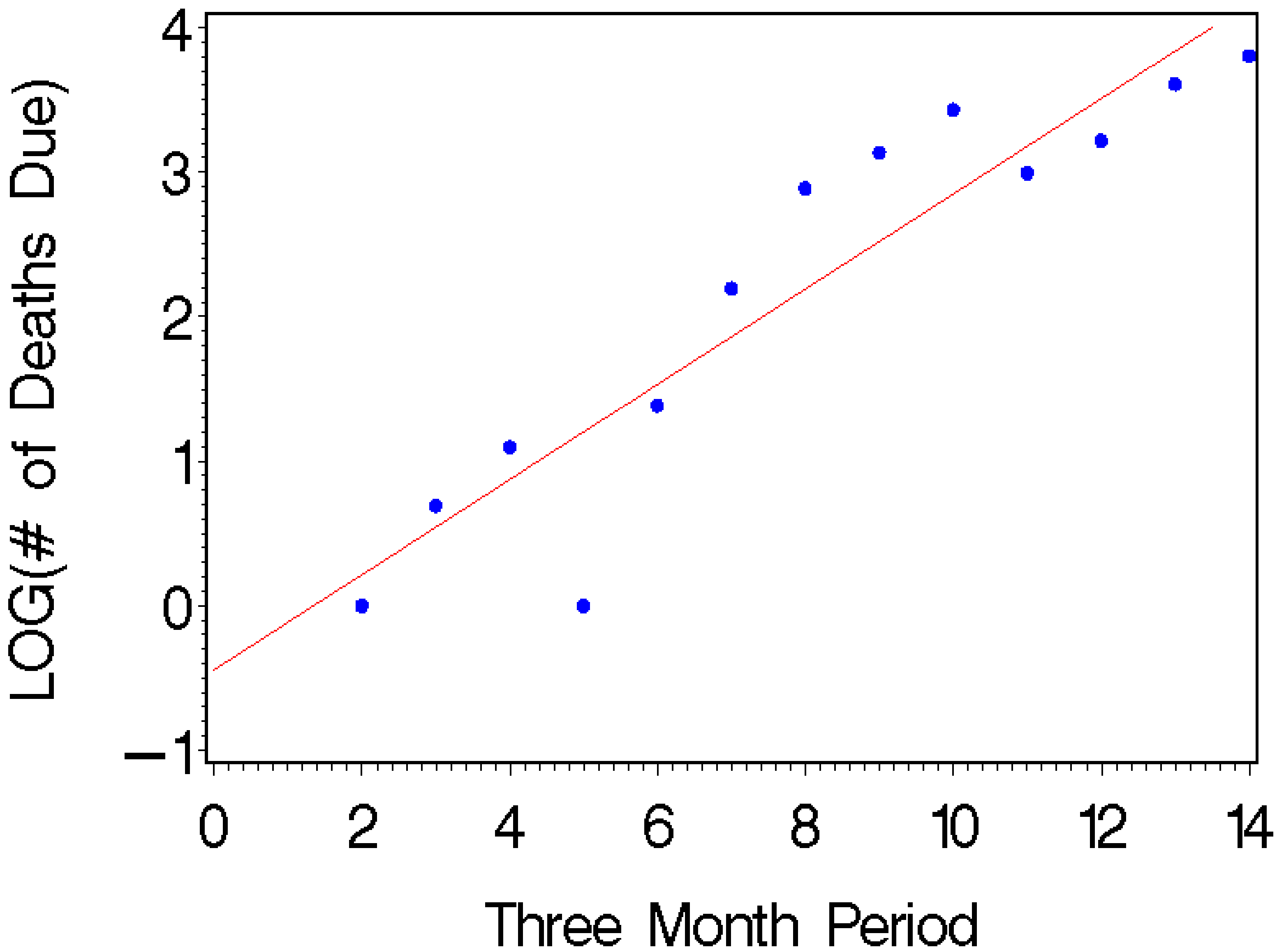
WARNING: The specified model did not converge.
ERROR: The mean parameter is either invalid or at a limit of its range for some observations.

What's wrong?...

A Look at the Bad Model (linear link)



Back to Data but Plot $\log(y_i)$ by Month



(line is linear regression line)

Poisson Log-Linear Model for Deaths

Figure suggests a log link might work better:

$$\log(\hat{\mu}_i) = .3396 + .2565 x_i$$

		$\hat{\mu}_i$ when Link is				$\hat{\mu}_i$ when Link is	
x_i	y_i	Log	Identity	x_i	y_i	Log	Identity
1	0	1.82	-4.21	8	18	10.93	12.69
2	1	2.35	-1.88	9	23	14.13	15.12
3	2	3.03	0.55	10	31	18.26	17.55
4	3	3.92	2.98	11	20	23.60	19.98
5	1	5.06	5.41	12	25	30.51	22.41
6	4	6.56	7.84	13	37	39.43	24.84
7	9	8.46	10.27	14	45	50.96	27.27

. . . and it looks like it fits much better.

Figure of Fitted log(count) from Log-linear

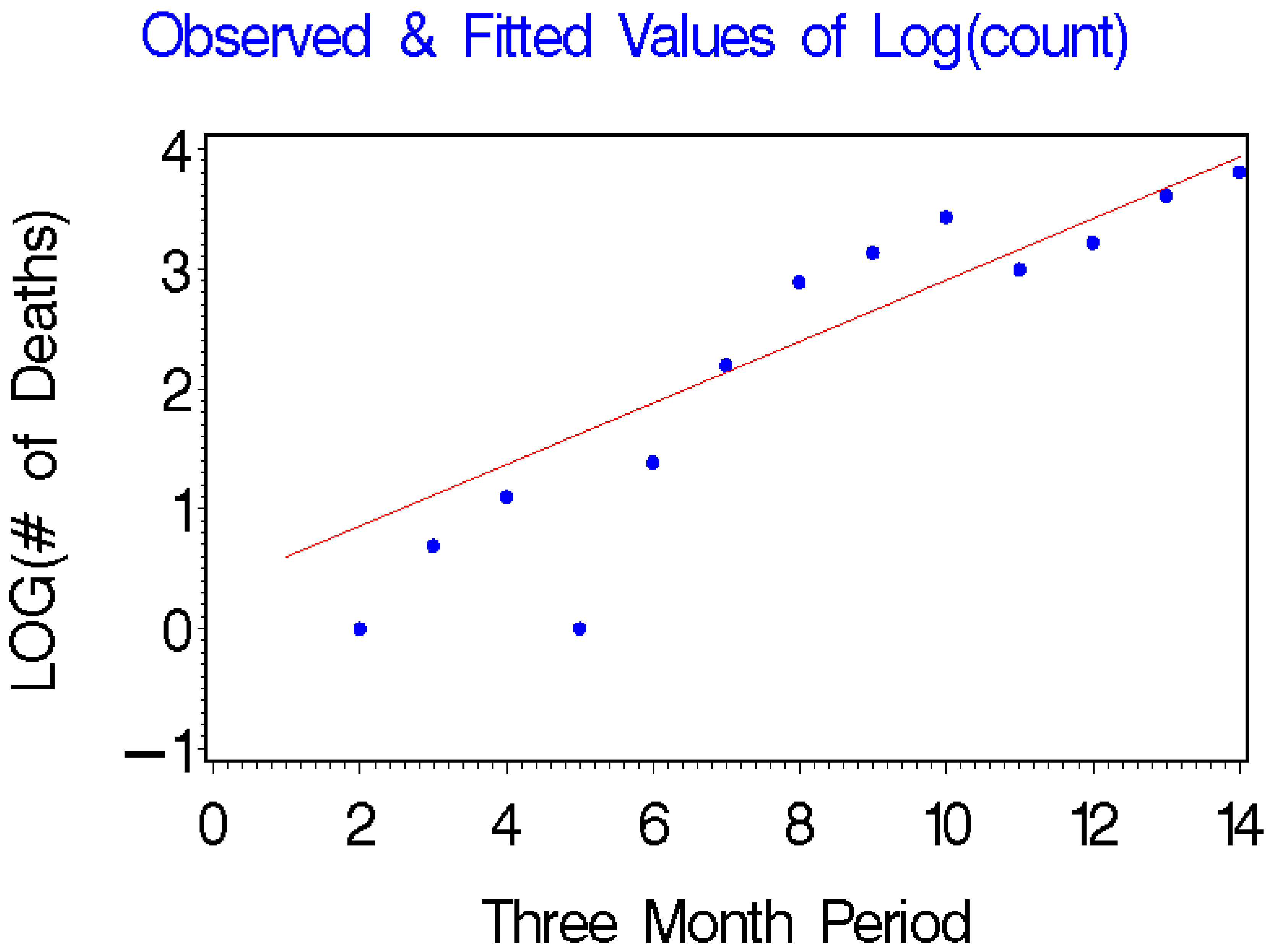
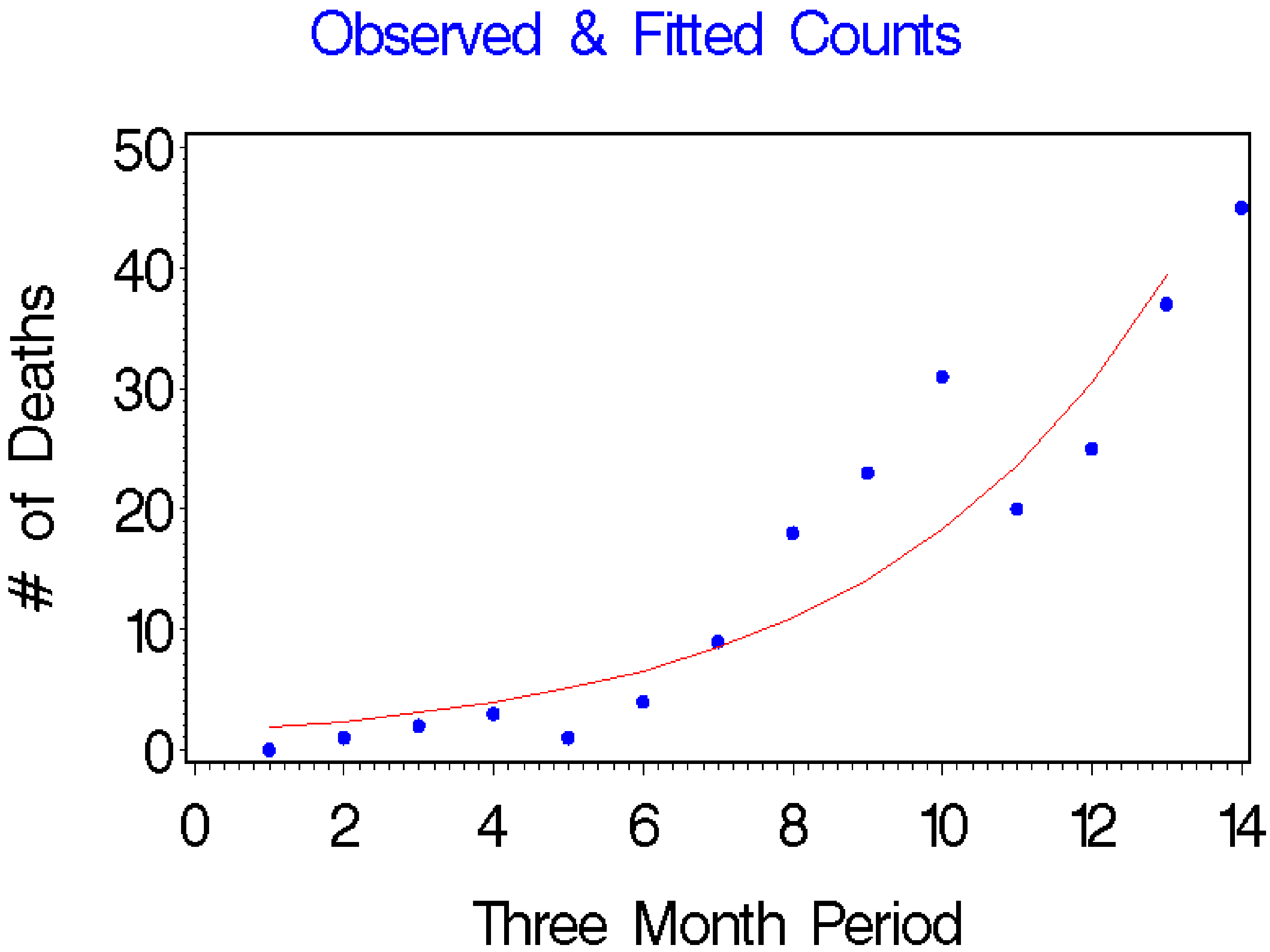


Figure of Fitted log(count) from Log-linear

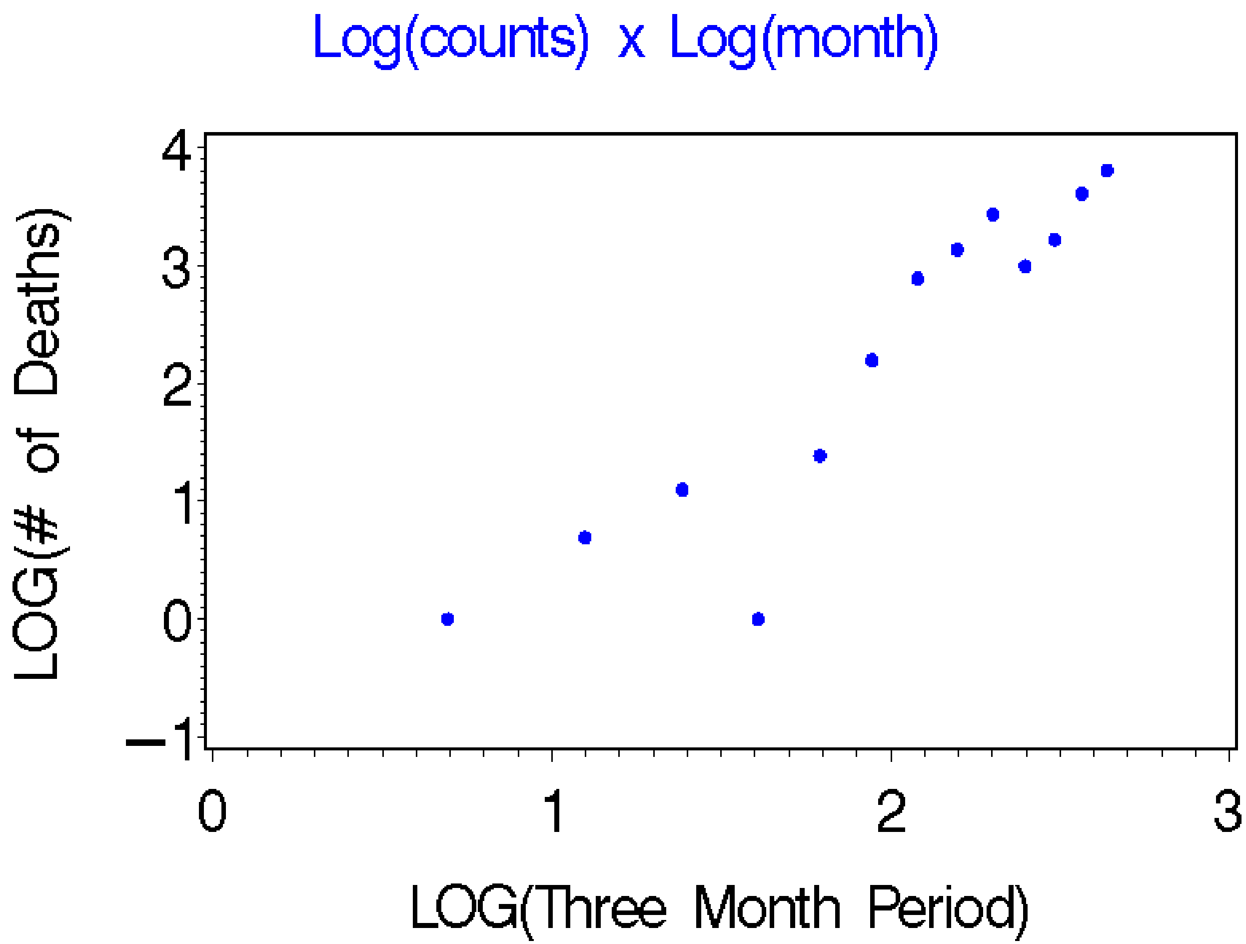


Pattern in residuals.

Transform explanatory variable

The number of deaths with low & high values of x_i are "over-fit" and number with middle x_i's are under-fit.

Transform $x_i \longrightarrow x_i^* = \log(x_i)$



Poisson Regression with Transformed x

The estimated GLM with model

- Random: Y follows Poisson distribution.
- Systematic: $\alpha + \beta \log(x_i) = \alpha + \beta x_i^*$
- Link: Log $\longrightarrow \log(\mu)$.

As a log-linear model

$$\log(\hat{\mu}_i) = -1.9442 + 2.1748 x_i^*$$

or equivalently, as a multiplicative model

$$\hat{\mu}_i = e^{-1.9442} e^{2.1748 x_i^*}$$

Interpretation: For a 1 unit increase in log(month), the estimated count increases by a factor of $e^{2.1748} = 8.80$

Is this "large"?

How Large is Large in a Statistical Sense?

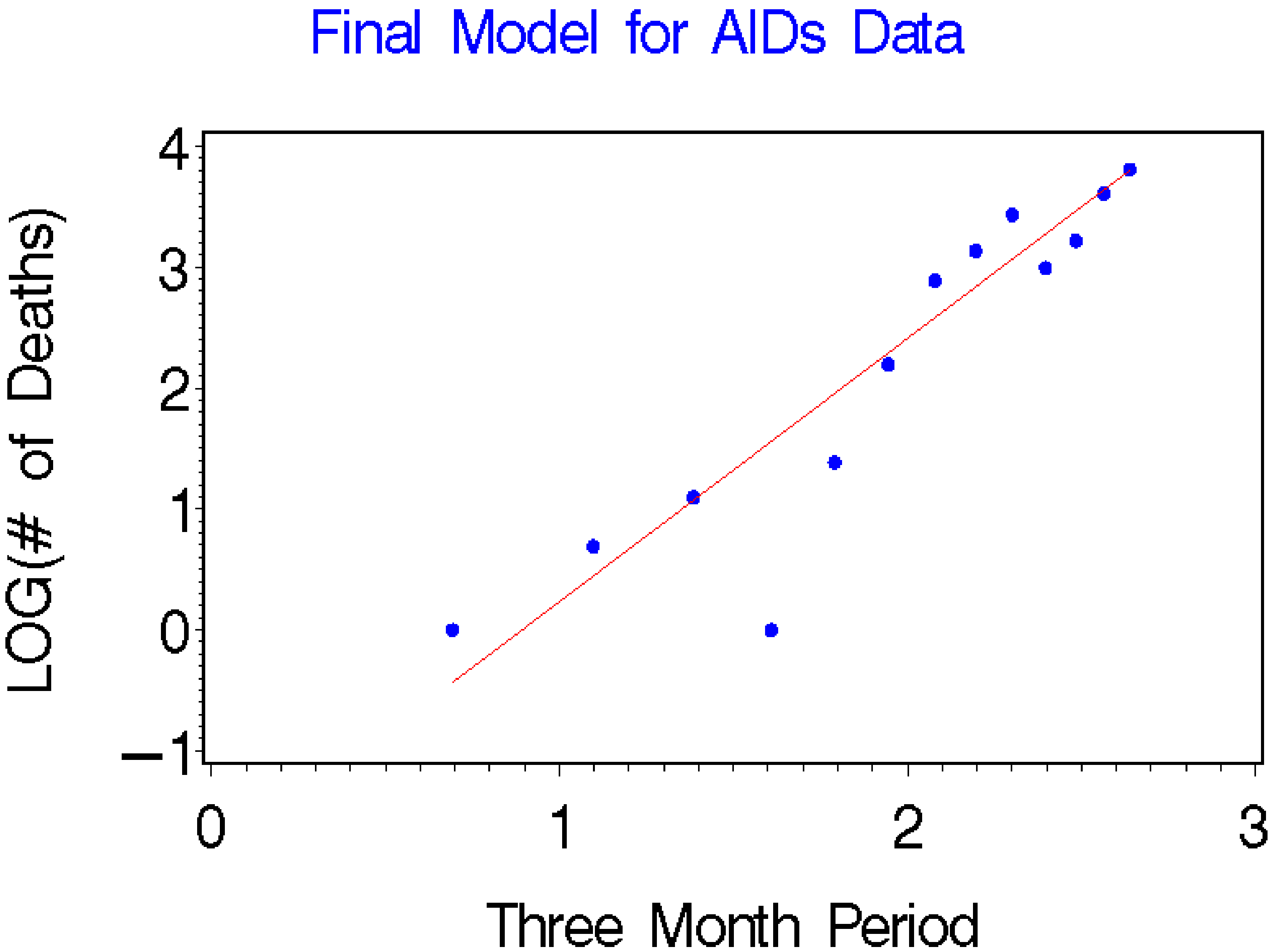
SAS/GENMOD provides asymptotic standard errors (ASE, i.e., large sample) for the parameter estimates.

The ASE for $\hat{\beta}$ equals $.2151$, and

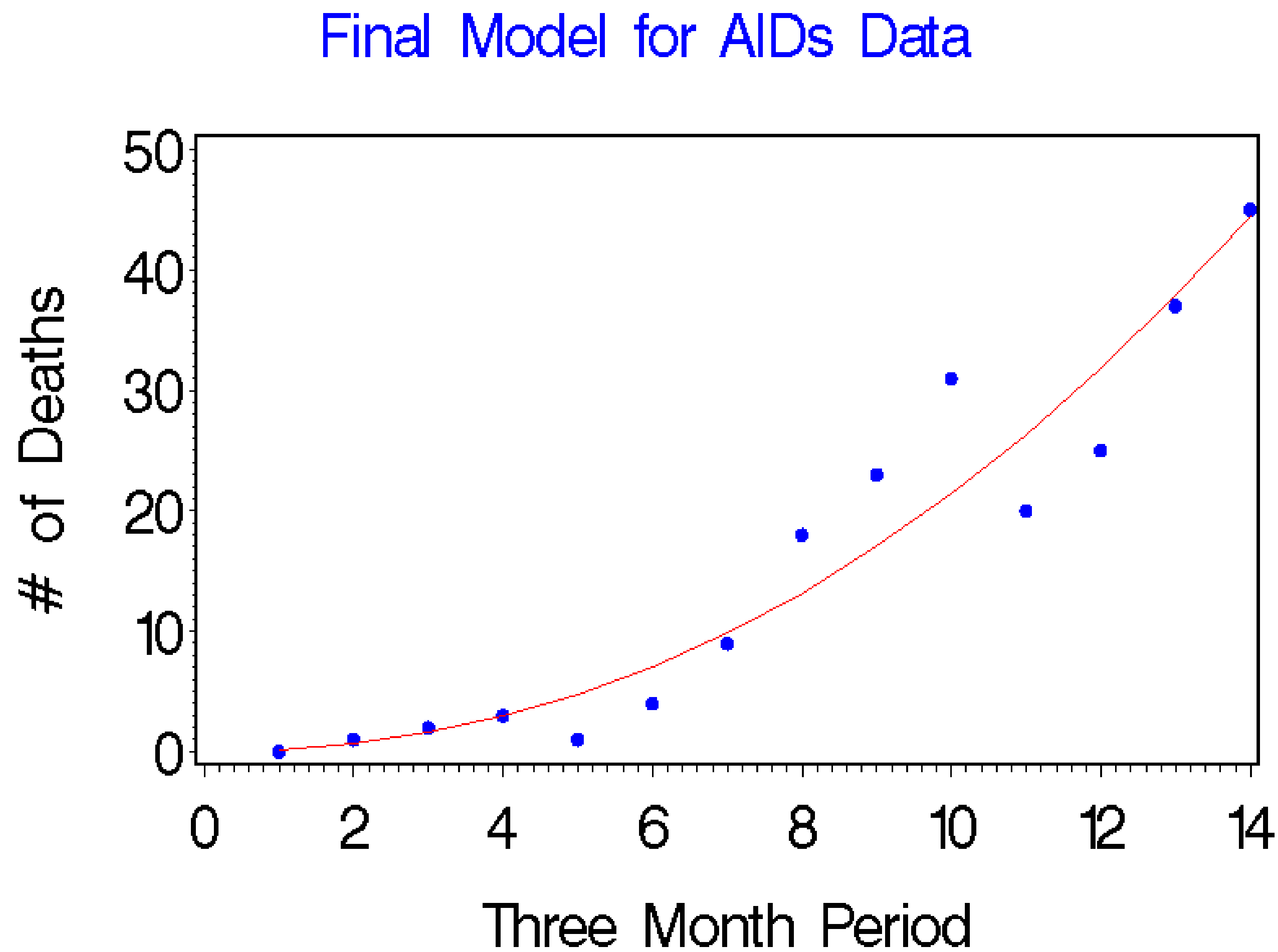
$$\hat{\beta} \pm 2(.2151) \longrightarrow (1.745, 2.605)$$

which suggests that this is large in a statistical sense.

Observed and Fitted Log(Counts)



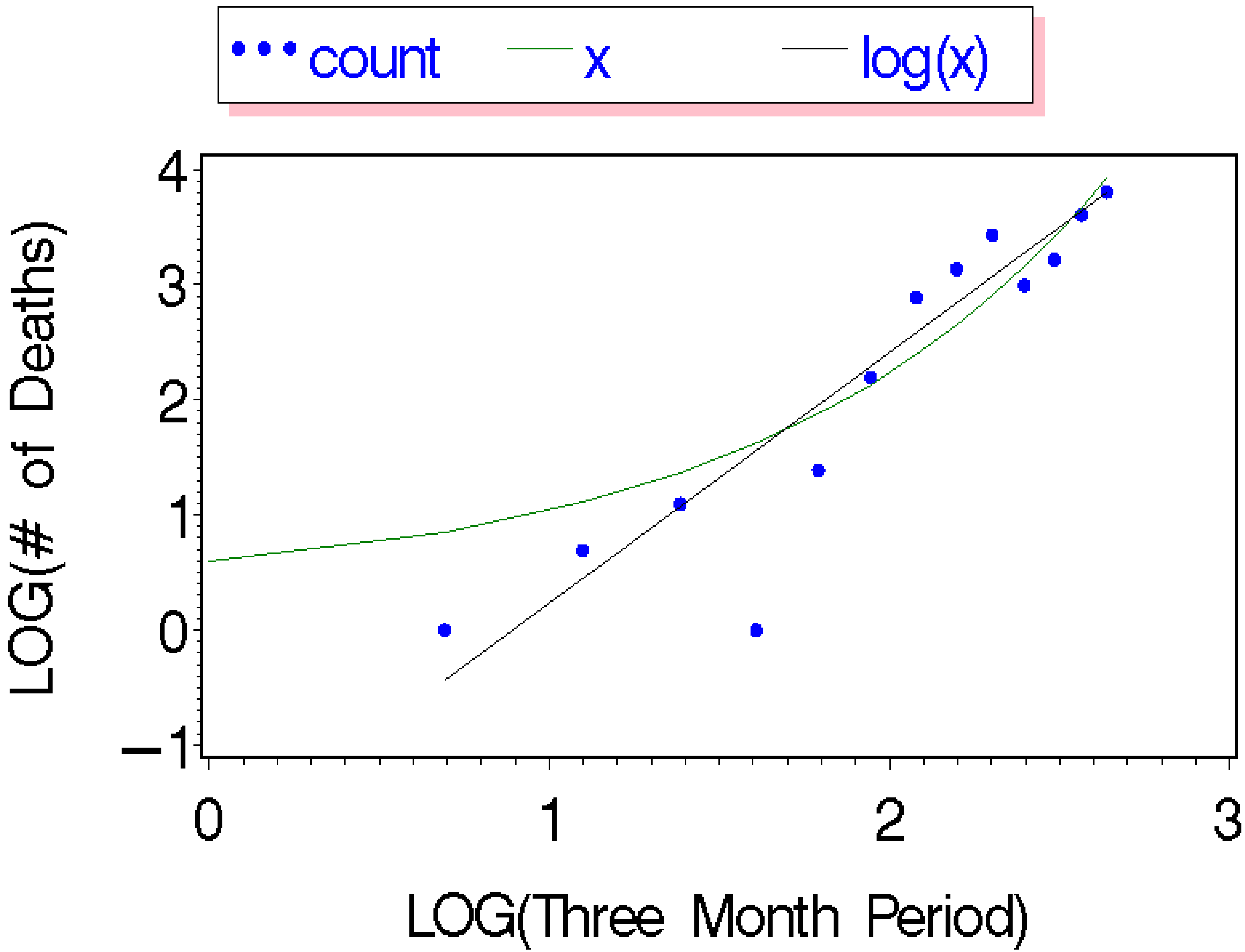
Observed and Fitted Counts



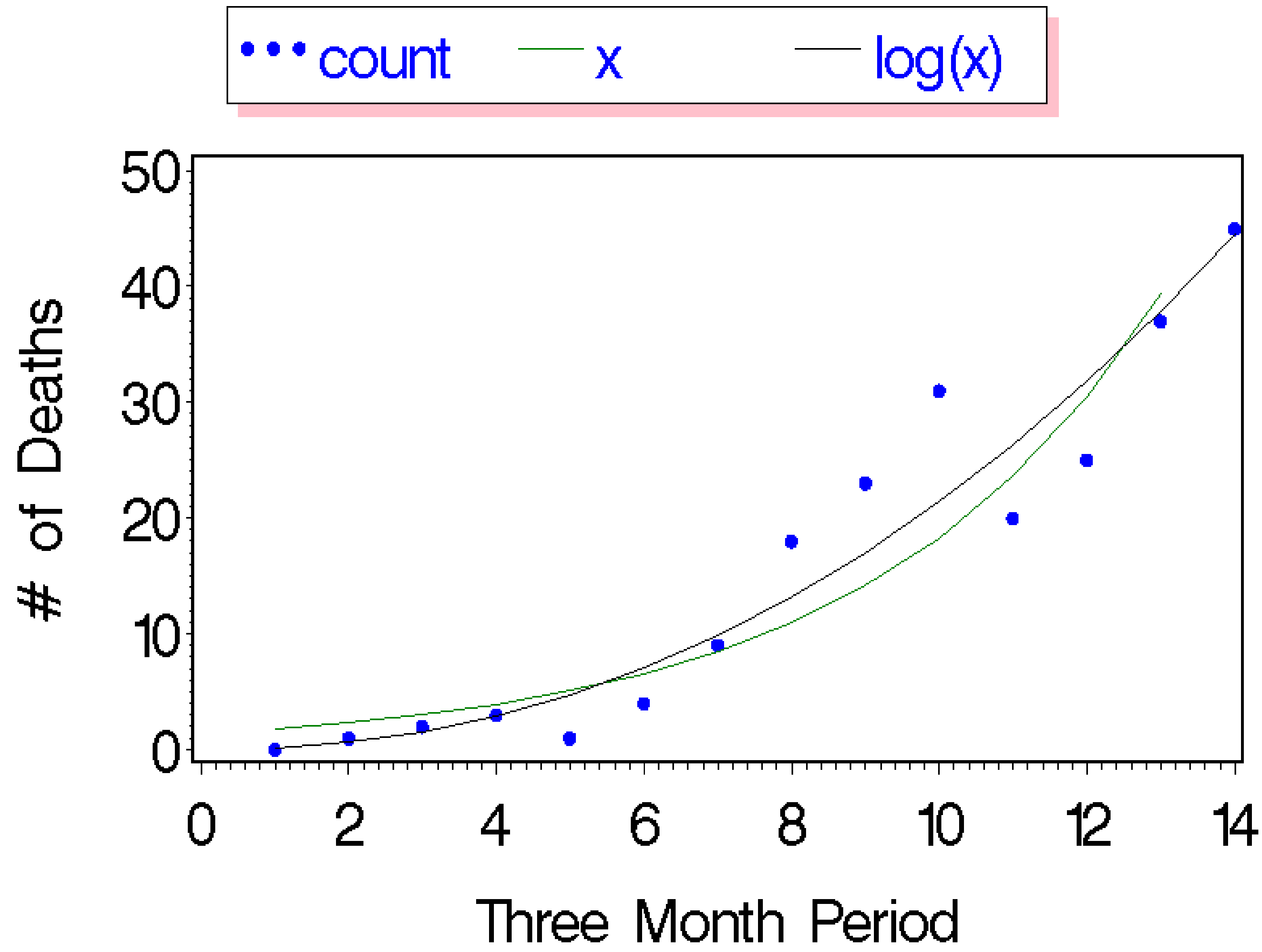
Comparison of Fitted Counts

		$\log(x_i)$	x_i	x_i
x_i	y_i	Log	Log	Identity
1	0	.14	1.82	-4.21
2	1	.65	2.35	-1.88
3	2	1.56	3.03	0.55
4	3	2.92	3.92	2.98
5	1	4.74	5.06	5.41
6	4	7.05	6.56	7.84
7	9	9.86	8.46	10.27
8	18	13.17	10.93	12.69
9	23	17.02	14.13	15.12
10	31	21.40	18.26	17.55
11	20	26.33	23.60	19.98
12	25	31.82	30.51	22.41
13	37	37.87	39.43	24.84
14	45	44.49	50.96	27.27

Comparison in Log-Scale



Observed and Fitted Counts



More Interpretation of Poisson Regression

- The marginal effect of x_i (month period) on μ_i (expected number of deaths due to AIDS).

  For a 1 unit increase in log(month), the estimated count increases by a factor of $e^{2.1748} = 8.80$.
- Computed fitted values and compared them to the observed. (table and plots of this).
- Additional one: We can look at the predicted probability of number of deaths given value on x_i. (This is not too useful here, but would be of use in a predictive setting).

Counts follow a Poisson distribution, so

$$P(Y_i = y) = \frac{e^{-\mu_i}\mu_i^y}{y!}$$

According to our estimated model, probabilities that the number of deaths equals y_i for particular value(s) of x_i is

$$P(Y_i = y) = \frac{e^{-e^{(-1.9442+2.1748x_i^*)}}e^{(-1.9442+2.1748x_i^*)^y}}{y!}$$

Probabilities of Number of Deaths

$$P(Y_i = y) = \frac{e^{-e^{(-1.9442+2.1748x_i^*)}} e^{(-1.9442+2.1748x_i^*)y}}{y!}$$

or since we already have $\hat{\mu}_i$ computed, we can use

$$P(Y_i = y) = \frac{e^{-\hat{\mu}_i} \hat{\mu}_i^y}{y!}$$

For example, consider quarter $= 3$ (and $\log(3) = 1.09861$), we have

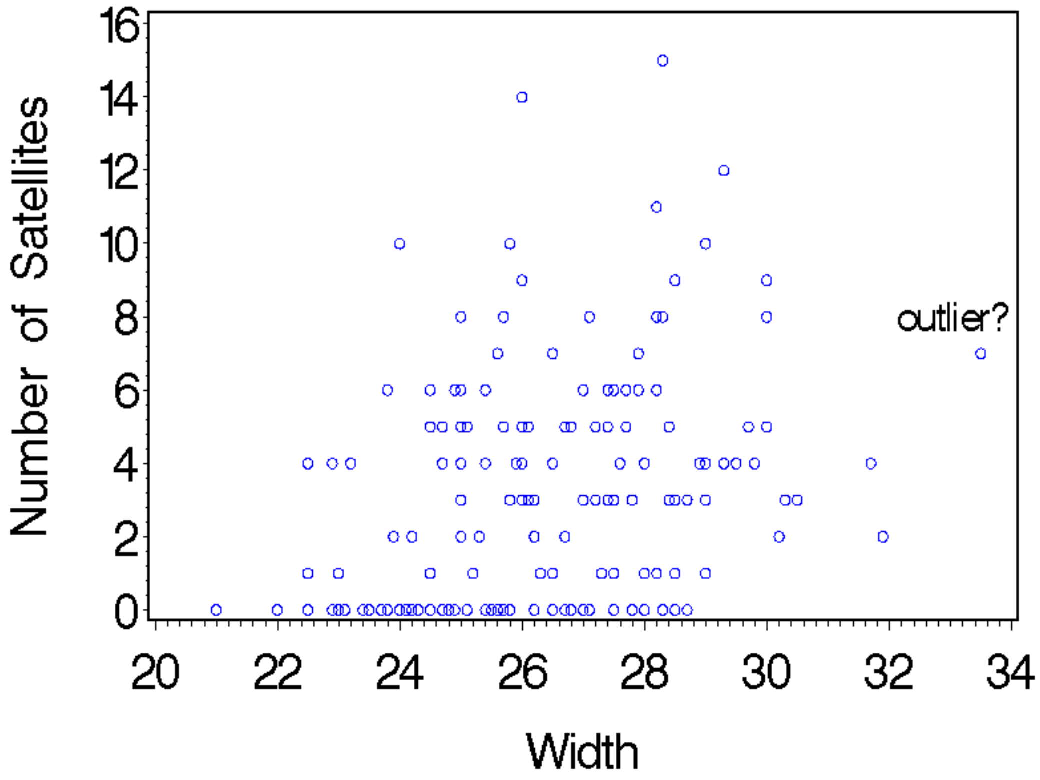
$$\hat{\mu}(\text{quarter} = 3) = 1.5606$$

$$\begin{aligned}
P(Y_3 = 0) &= e^{-1.5606}(1.5606)^0/0! = .210 \\
P(Y_3 = 1) &= e^{-1.5606}(1.5606)^1/1! = .328 \\
P(Y_3 = 2) &= e^{-1.5606}(1.5606)^2/2! = .128 \\
&\vdots \\
P(Y_3 = 10) &= e^{-1.5606}(1.5606)^{10}/10! = .000000253
\end{aligned}$$

Example 2: Crab Data

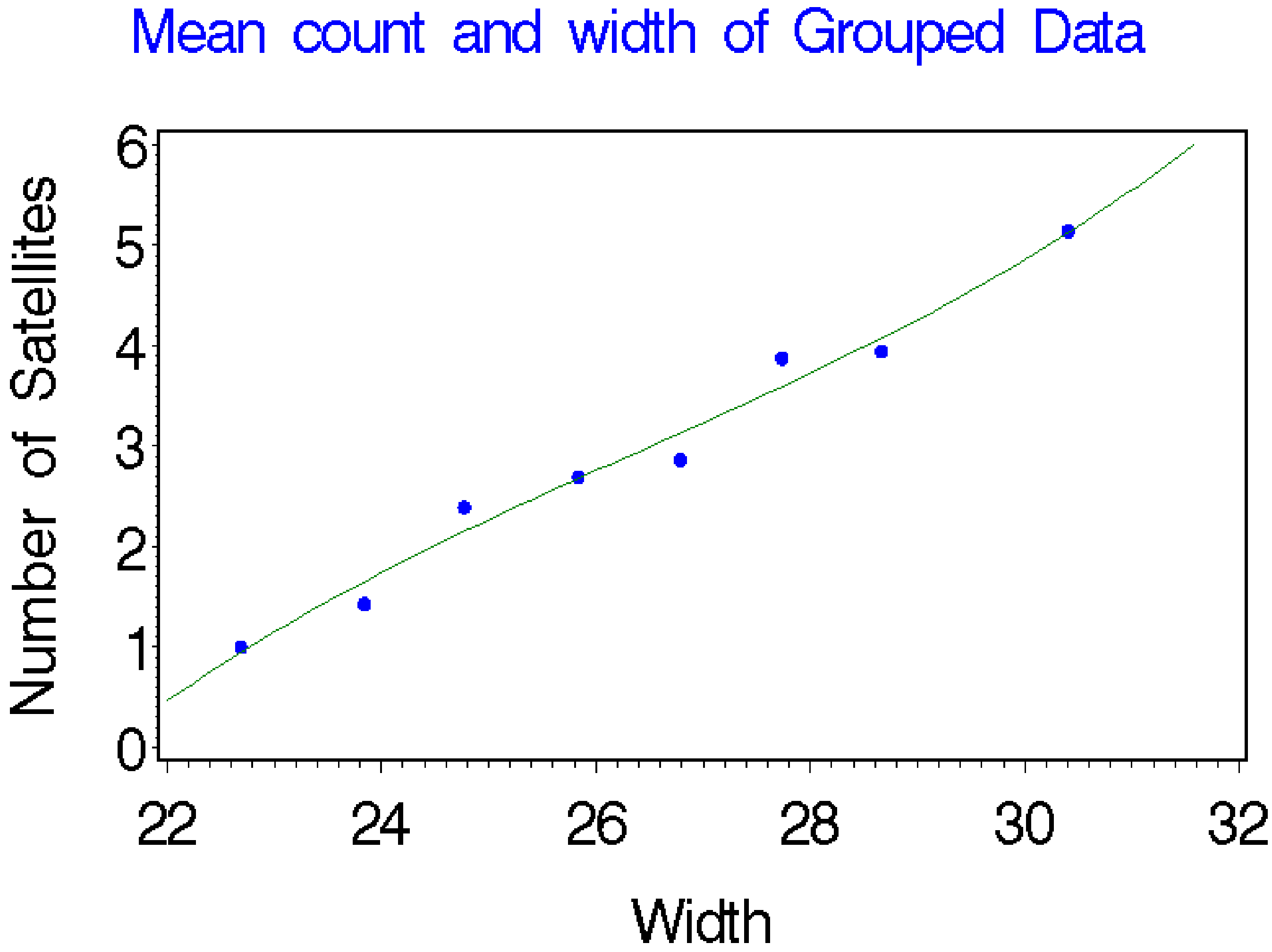
Agresti (1996)'s horseshoe crab data.

- Response variable is the number of satellites a female horseshoe crab has (i.e., how many males are attached to her).
- Explanatory variable is the width of the female's back.



A Smoother Look

The data were collpased into 8 groups by their width (i.e., ≤ 23.25, 23.25–24.25, 24.25–$25.25\ldots$, > 29.25).



Estimated Poisson Regression for Crabs

$$\log(\hat{\mu}_i) = -3.3048 + .1640x_i$$

- The estimated ASE of $\hat{\beta} = .164$ equals .020, which is small relative to $\hat{\beta}$.
- Since $\hat{\beta} > 0$, the wider the female crab, the greater the expected number of satellites. Note: $\exp(.1640) = 1.18$.
- There is an outlier (with respect to the explanatory variable).
  - Question: how much does this outlier effect the fit of the model?
  - Answer: Remove it and re-estimate the model.

$$\log(\hat{\mu}_i) = -3.4610 + .1700x_i$$

  and ASE of $\hat{\beta} = .1700$ equals $.0216$.
  - So in this case, it doesn't have much effect. . .
    The same basic result holds (i.e., positive effect of width on number of satellites, $\hat{\beta}$ is "significant" and similar in value).

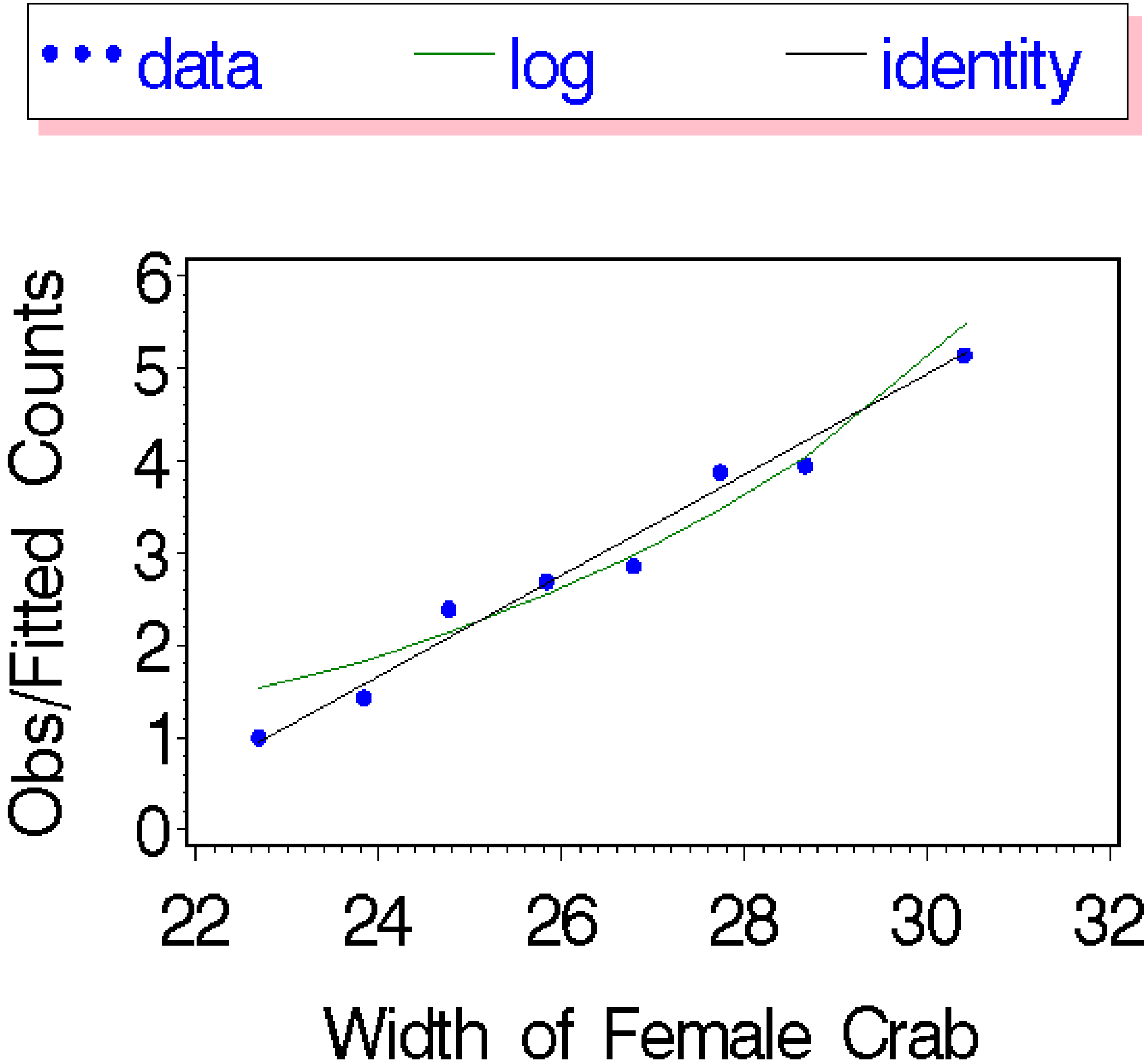
Poisson Regression with Identity Link

- From the figure of collapsed data, it looks like either a linear or a log link might work.
- The estimated model with the linear link :

$$\hat{\mu}_i = -11.53 + .55x_i$$

- Since the effect on the number of expected satellites of female width (μ_i) is linear and $\hat{\beta} = .55 > 0$, as width increases by 1 cm, the expected count increases by .55.
- Question: Is the Poisson regression model with the linear or the logit link better for these data?
- Answer: Quick look but more formal later when we discuss model assessment (or read further in the text).

Log versus Identity Link for Crabs



Poisson regression for rates

Events occur over time (or space), and the length of time (or amount of space) can vary from observation to observation. Our model should take this into account.

Example: Gardner, Mulvey, & Shaw (1995), *Psychological Bulletin, 118*, 392–404.

$Y =$ Number of violent incidents exhibited over a 6 month period by patients who had been treated in the ER of a psychiatric hospital.

During the 6 months period of the study, the individuals were primarily residing in the community. The number of violent acts depends on the opportunity to commit them; that is, the number of days out of the 6 month period in which a patient is in the community (as opposed to being locked up in a jail or hospital).

The Data

put histogram here.

Poisson Regression for Rates of Events

$Y =$ count (e.g., number violent acts).
$t =$ index of the time or space (e.g., days in the community).

The sample rate of occurrence is Y/t.
The expected value of the rate is

$$E(Y/t) = \frac{1}{t}E(Y) = \mu/t$$

The Poisson loglinear regression model for the expected rate of the occurrence of events is

$$\begin{aligned} \log(\mu/t) &= \alpha + \beta x \\ \log(\mu) - \log(t) &= \alpha + \beta x \\ \log(\mu) &= \alpha + \beta x + \log(t) \end{aligned}$$

The term "$-\log(t)$" is an adjustment term and each individual may have a different value of t.

The term $-\log(t)$ is referred to as an "offset".

As a Multiplicative Model

The Poisson loglinear regression model with a log link for rate data is

$$\begin{aligned} \log(\mu/t) &= \alpha + \beta x \\ \mu/t &= e^{\alpha} e^{\beta x} \\ \mu &= t e^{\alpha} e^{\beta x} \end{aligned}$$

The expected value of counts depends on both t and x, both of which are observations (i.e., neither is a parameter of the model).

Gardner, Mulvey, & Shaw (1995)

- Response variable is rate of violent incidents, which equals the number of violent incident divided by the number of days an individual resided in the community. ($\bar{y} = 3.0$ with $s = 7.3$ and $\bar{t} = 154$ with $s = 42$ days).
- Explanatory variables:
  - Age ($\bar{x}_1 = 28.6$ years and $s_1 = 11.1$)
  - Sum of 2 ER clinicans ratings of concern on a 0 – 5 scale, so x_2 ranges from 0 to 10. ($\bar{x}_2 = 2.9$ with $s_2 = 3.1$).
  - History of previous violent acts, where

$$\begin{aligned} x_3 &= 0 \quad \text{means no previous acts} \\ &= 1 \quad \text{previous act either 3 days before } \textit{or} \\ &\qquad \text{more than 3 days before} \\ &= 2 \quad \text{previous acts both 3 days before } \textit{and} \\ &\qquad \text{more than 3 days before} \end{aligned}$$

r(concern, history)$= .55$
r(age,history)$= -.11$
r(age,concern)$= -.07$

Estimated Parameters

Coefficient	Value	ASE	value/ASE
Intercept	-3.410	.0690	-49.29
Age	-.045	.0023	-19.69
Concern	.083	.0075	11.20
History	.420	.0380	11.26

put histogram of fitted and observed

Note: Poisson regression models for rate data are related to models for “survival times”.

SAS/PROC GENMOD


```
DATA violent;
    INPUT age history concern days count;
    Ldays = LOG(days);
```

To fit model for rate data with “days” as t:

```
PROC GENMOD DATA=violent;
    MODEL count = age history concern
    / LINK=log DIST=poisson OFFSET=Ldays;
```

To include interactions, change model statement

```
MODEL count = age history concern history*concern
      / LINK=log DIST=poisson OFFSET=Ldays;
```

If we want to treat “history” as a discrete (nominal) variable, you would include the statement

```
CLASS history;
```

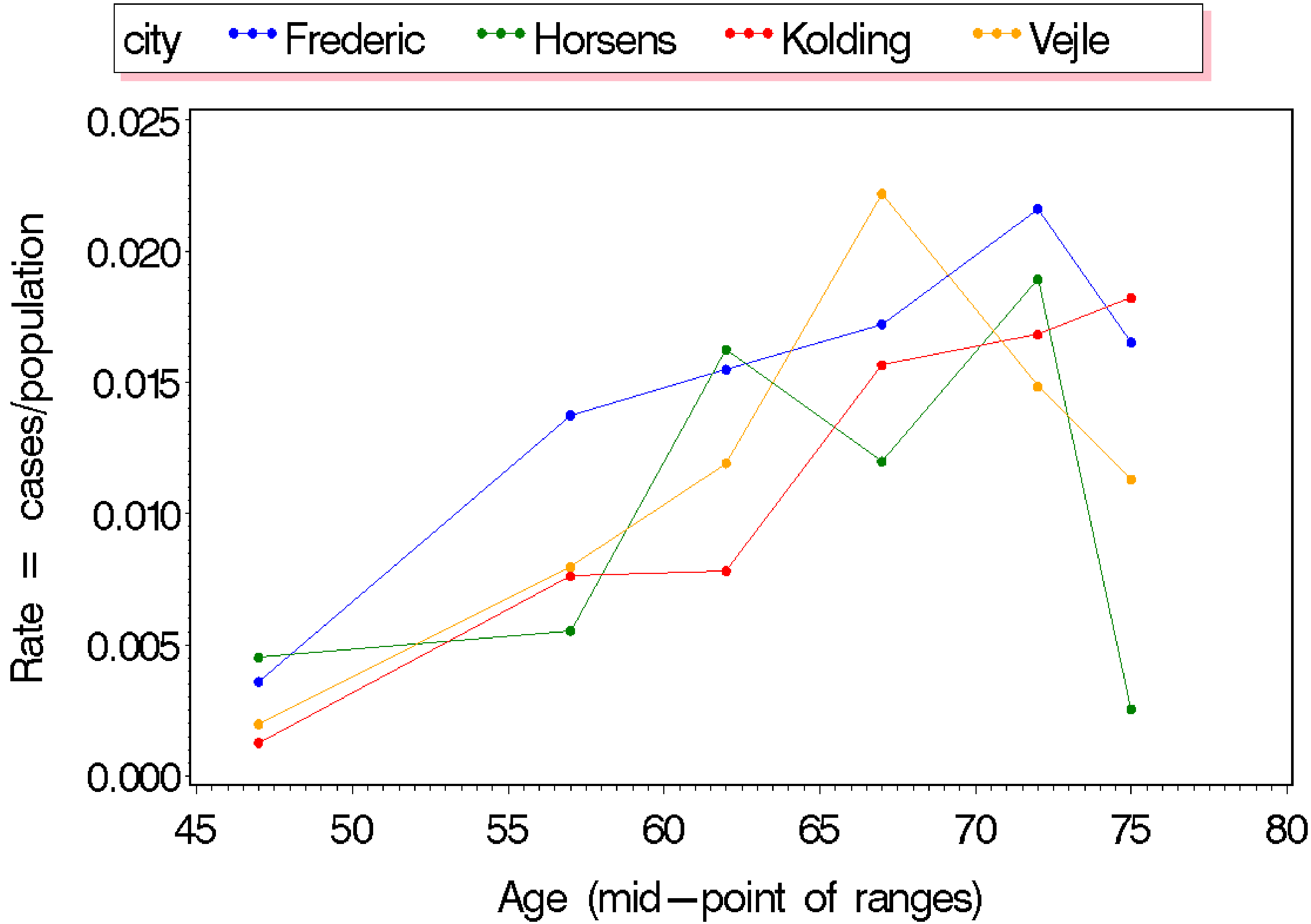
If the model with “history” treated as a numerical variable almost as well as the model with “history” treated as a nominal discrete variable, which model would you prefer?

Example 2: Lung Cancer

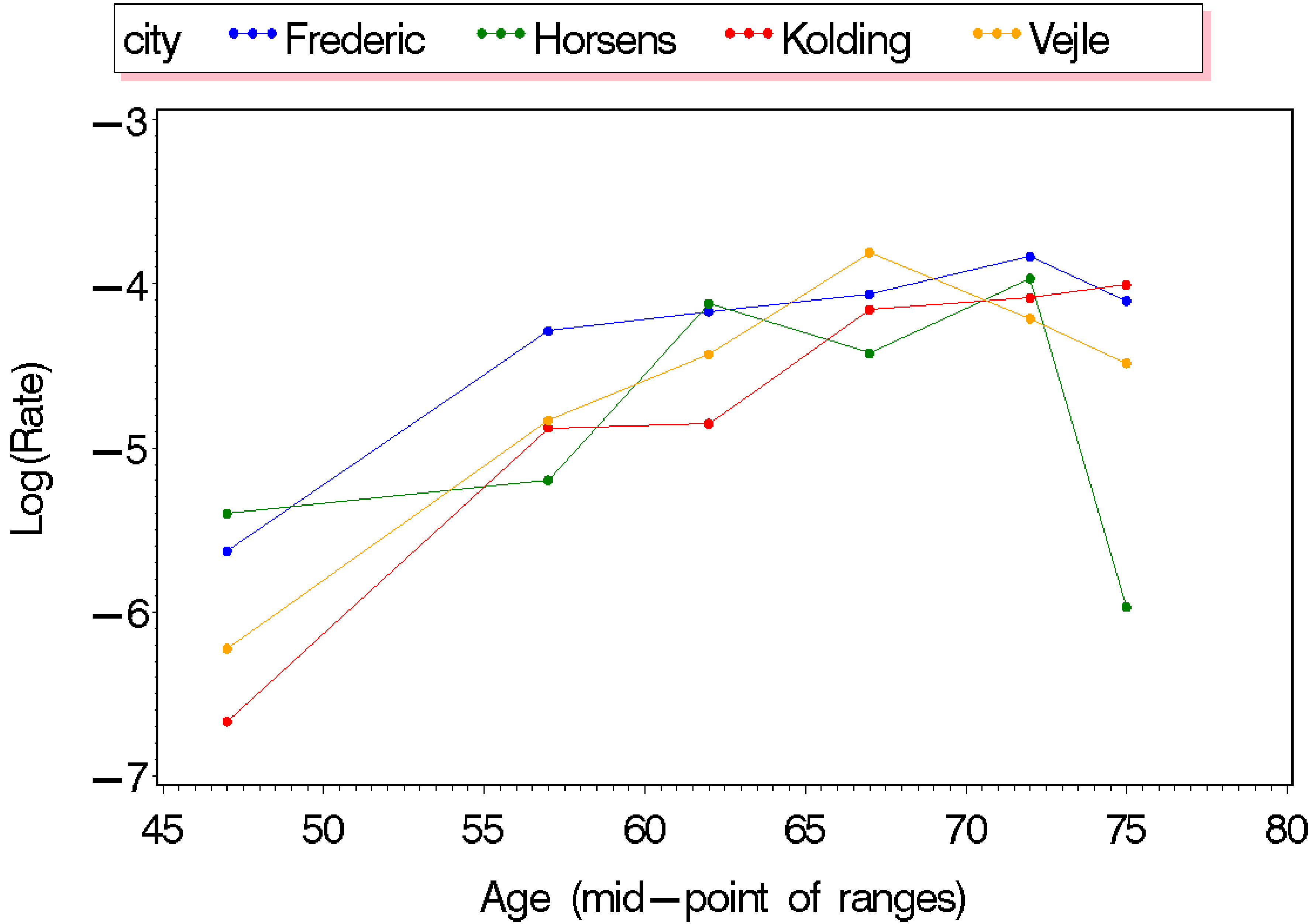
Data are from Lindsey (1995) from Andersen (1977)

- Response Variable: $Y =$ Number of cases of lung cancer and it follows a Poission distribution.:
- Explanatory Variables:
  - City in Denmark (Fredericia, Horsens, Kolding, Vejle).
  - Age (40–54, 55–59, 60–64, 65–69, 70–74, $>$75).
- Offset $=$ Population size of each age group of each city.
- We will model the rate of cases of lung cancer $= Y/t$.

Plot of the Rate by Age



Plot of the log(Rate) by Age



Model 1: Age and City both Nominal

Define

$$\text{Fredericia} = \begin{cases} 1 & \text{if city is Frederica} \\ 0 & \text{other city} \end{cases}$$

$$\text{Horsens} = \begin{cases} 1 & \text{if city is Horsens} \\ 0 & \text{other city} \end{cases}$$

$$\text{Kolding} = \begin{cases} 1 & \text{if city is Kolding} \\ 0 & \text{other city} \end{cases}$$

Define Dummy variables for the 6 age classes (groups).

Model 1:

$$\begin{aligned} \log(Y/\text{pop}) &= \alpha + \beta_1(\text{Fredericia}) + \beta_2(\text{Horsens}) + \beta_3(\text{Kolding}) \\ &= \beta_4(\text{Age1}) + \beta_5(\text{Age2}) + \beta_6(\text{Age3}) + \beta_7(\text{Age4}) \\ &\quad \beta_8(\text{Age5}) \end{aligned}$$

Parameter Estimates from Model 1

Parameter			Estimate	df	s.e.	X^2	p
Intercept		α	1	-4.48	0.21	423.33	$< .01$
city	Frederic	β_1	1	0.27	0.18	2.10	.15
city	Horsens	β_2	1	-0.05	0.19	0.09	.76
city	Kolding	β_3	1	-0.09	0.19	0.25	.62
city	Vejle		0	0.00	0.00	.	.
age	40-54	β_4	1	-1.41	0.25	32.18	$< .01$
age	55-59	β_5	1	-0.31	0.25	1.60	.21
age	60-64	β_6	1	0.09	0.23	0.18	.67
age	65-69	β_7	1	0.34	0.23	2.22	.14
age	70-74	β_8	1	0.43	0.23	3.34	.07
age	>75		0	0.00	0.00	.	.

Note: $G^2 = 23.45$, $df = 15$, $p = .08$

Model 2: City Nominal & Age Numerical

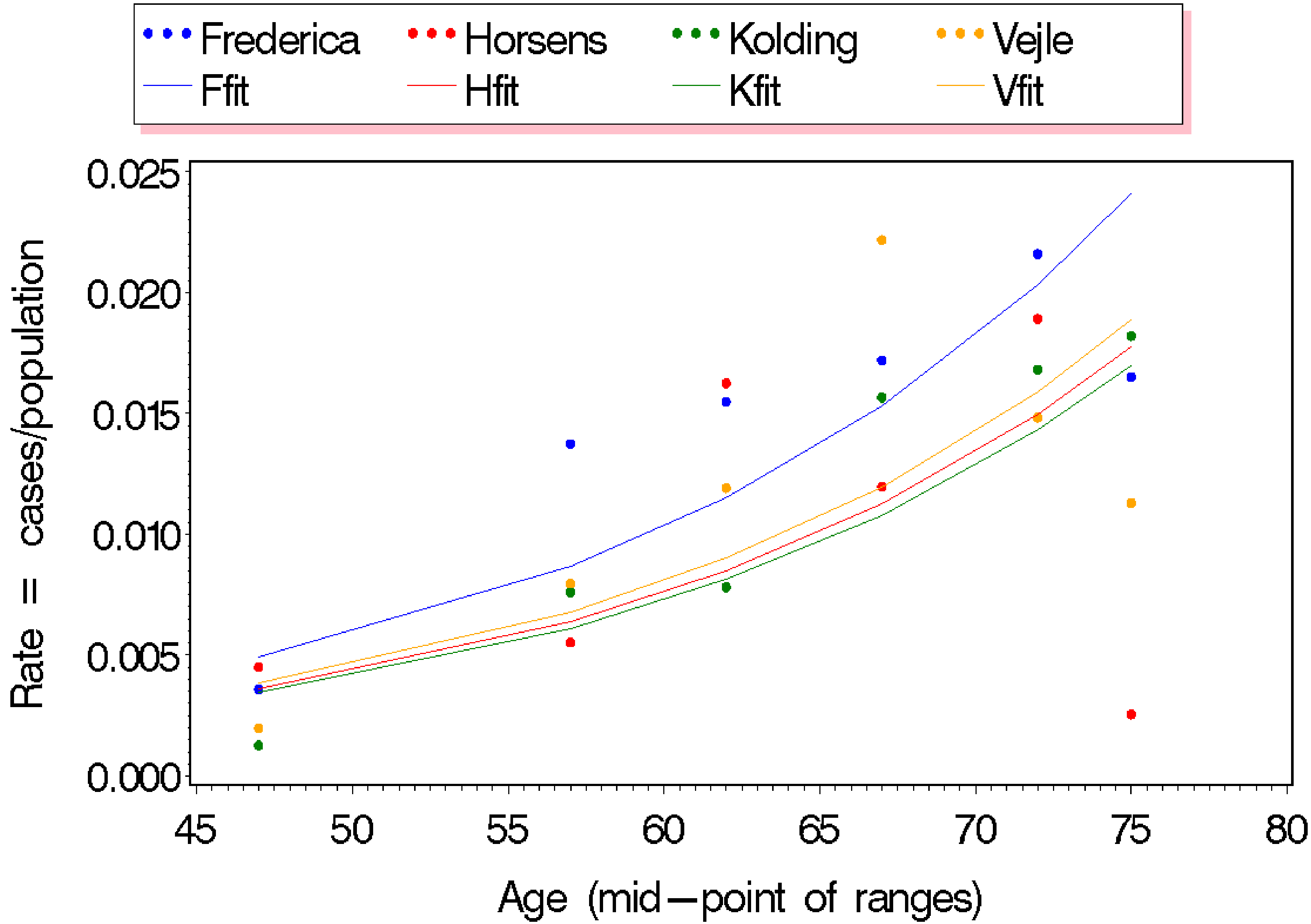
The mid-point of the age ranges were used (except for the last one, I used 75).

$$\begin{aligned}\log(Y/\text{pop}) &= \alpha + \beta_1(\text{Fredericia}) + \beta_2(\text{Horsens}) + \beta_3(\text{Kolding}) \\ &= \beta_4(\text{Age Mid-point})\end{aligned}$$

Parameter			Estimate	df	s.e.	X^2	p
Intercept		α	1	-8.22	0.44	349.18	$< .01$
city	Frederic	β_1	1	0.24	0.18	1.72	0.19
city	Horsens	β_2	1	-0.05	0.19	0.10	0.76
city	Kolding	β_3	1	-0.10	0.19	0.28	0.60
city	Vejle		0	0.00	0.00	.	.
age-midpoint		β_4	1	0.05	0.00	75.62	$< .01$

Note: $G^2 = 46.45$, $df = 19$, $p < .01$

Model 2: Observed and Fitted Values



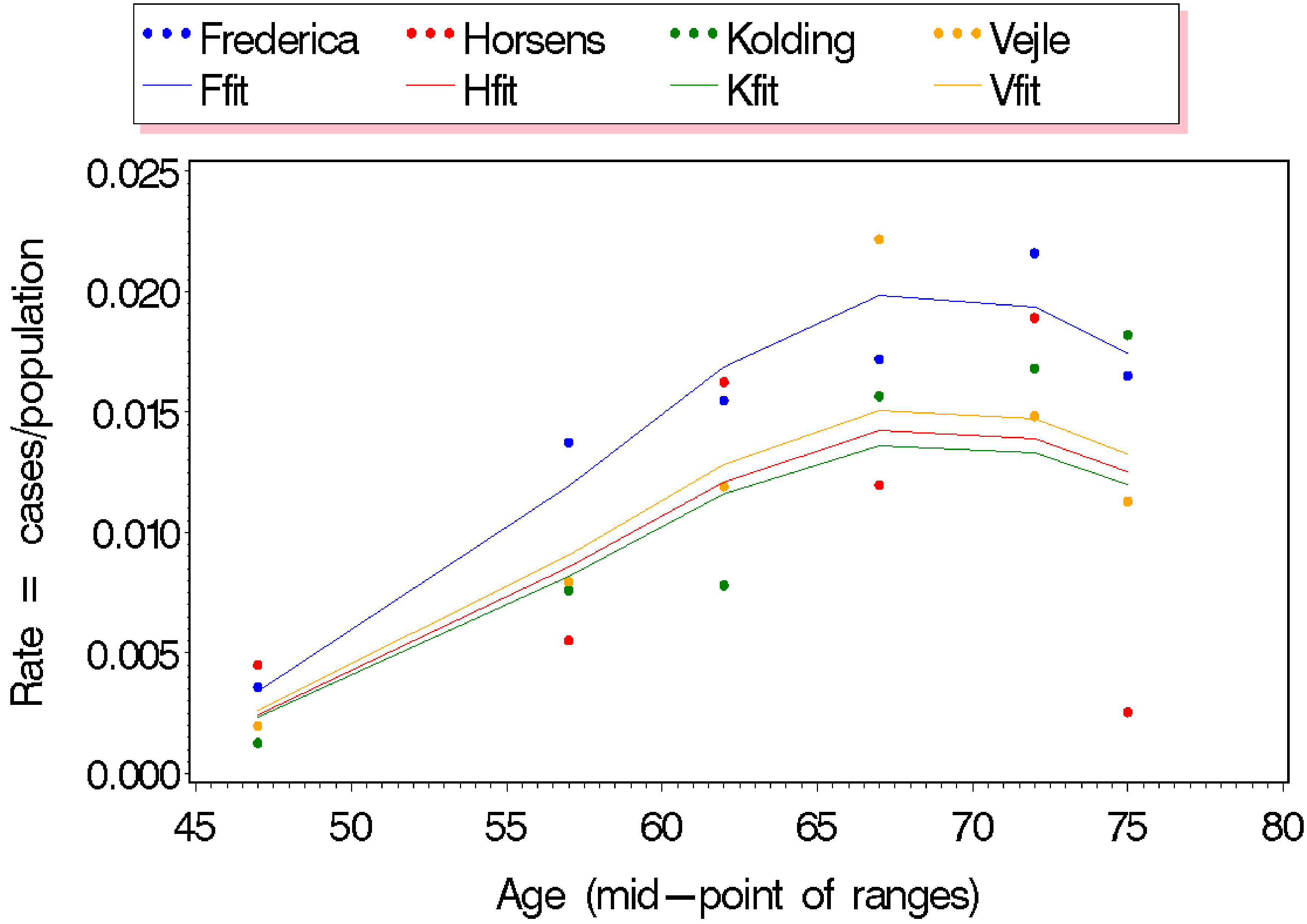
Model 3: City Nominal & Age Quadratic

$$
\begin{aligned}
\log(Y/\text{pop}) &= \alpha + \beta_1(\text{Fredericia}) + \beta_2(\text{Horsens}) + \beta_3(\text{Kolding}) \\
&= \beta_4(\text{Age Mid-point}) + \beta_5(\text{Age Mid-point})^2
\end{aligned}
$$

Parameter			Estimate	df	s.e.	X^2	p
Intercept		$\alpha 1$	1	-21.72	3.09	49.24	$< .01$
city	Frederic	β_1	1	0.27	0.18	2.13	0.14
city	Horsens	β_2	1	-0.05	0.19	0.09	0.76
city	Kolding	β_3	1	-0.10	0.19	0.26	0.61
city	Vejle		0	0.00	0.00	.	.
age-midpoint		$\beta_4 1$	1	0.50	0.10	24.91	$< .01$
age^2		$\beta_5 1$	1	-0.00	0.00	19.90	$< .01$

$G^2 = 26.02$, $df = 18$, $p = .10$.

Model 3: Observed and Fitted Values



Model 4: Simpler city & Age Quadratic

$$\text{Define Fredericia} = \begin{cases} 1 & \text{if city is Frederica} \\ 0 & \text{other city} \end{cases}$$

$$\log(Y/\text{pop}) = \alpha + \beta_1(\text{Fredericia}) + \beta_2(\text{Age Mid-point}) + \beta_3(\text{Age Mid-point})^2$$

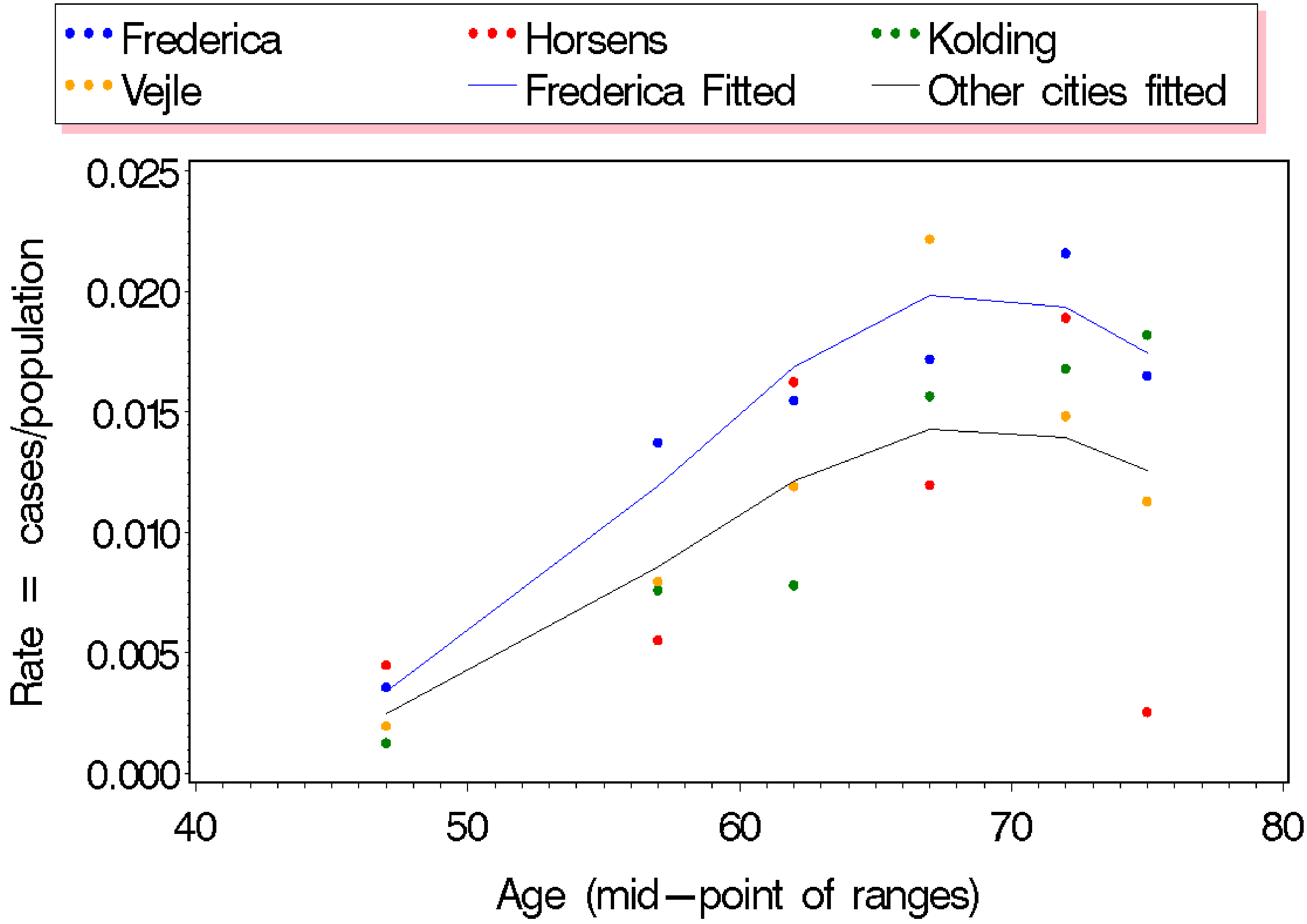
That is,

$$\log(Y/\text{pop}) = \begin{cases} \alpha + \beta_1 + \beta_2(\text{Age}) + \beta_3(\text{Age})^2 & \text{if Fredericia} \\ \alpha + \beta_2(\text{Age}) + \beta_3(\text{Age})^2 & \text{if other city} \end{cases}$$

Parameter			*df*	Estimate	s.e.	X^2	p
Intercept		$\alpha 1$	1	-21.78	3.09	49.61	$< .01$
frederic	1	β_1	1	0.32	0.14	4.92	.03
frederic	0		0	0.00	0.00	.	.
age-midpoint		β_2	1	0.50	0.10	24.93	$< .01$
age^2		β_3	1	-0.00	0.00	19.91	$< .01$

Note: $G^2 = 26.2815$, $df = 20$, $p = .16$.

Model 4: Fitted and Observed



Next Steps

Statistical Inference for Poisson Regression. . .