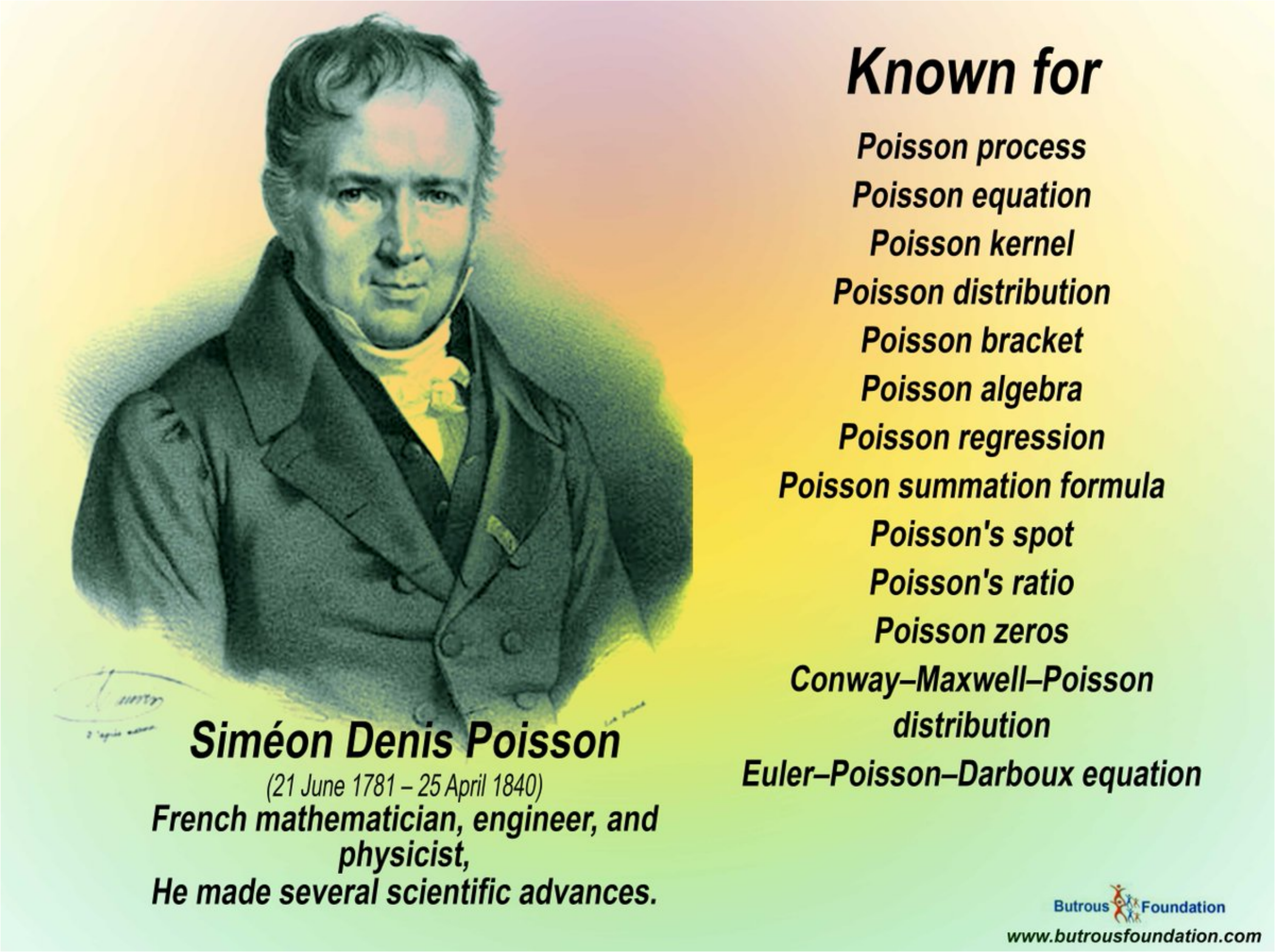
BST 210
Applied Regression Analysis



Known for

Poisson process
Poisson equation
Poisson kernel
Poisson distribution
Poisson bracket
Poisson algebra
Poisson regression
Poisson summation formula
Poisson's spot
Poisson's ratio
Poisson zeros
Conway–Maxwell–Poisson distribution
Euler–Poisson–Darboux equation

Siméon Denis Poisson

(21 June 1781 – 25 April 1840)

French mathematician, engineer, and physicist,
He made several scientific advances.

Butrous Foundation
www.butrousfoundation.com

Lecture 21
Plan for Today

- Poisson data
  - Review Y ~ Poisson(λ*t*)
  - Real world examples (will add to slides after class)
  - Poisson regression modeling
  - IRR interpretation
  - GOF tests

- Where does Poisson model break down?
  - Poisson model extensions:
    - Zero-Inflated Poisson (ZIP)
    - Negative Binomial

Recall: Poisson Distribution

- Recall discrete Poisson random variable Y ~ Poisson(λt), defined by

$$P(Y = y) = \frac{e^{-\lambda t} (\lambda t)^{y}}{y!}$$

where

Y counts number of events occurring in space/time interval

y = 0, 1, 2, … are independent

E(Y) = Var(Y)

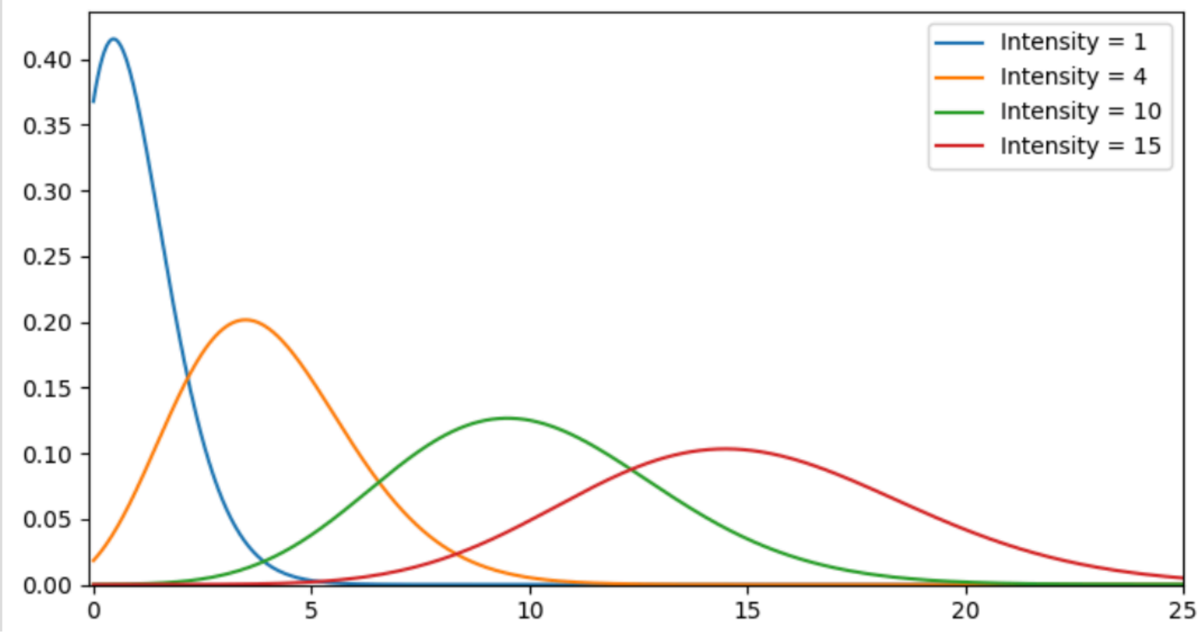
λ = (constant) number of cases per unit time (incidence rate)

t = person-time or area of exposure (non-overlapping)

Y are rare events (many trials; smaller probability of success)

Recall: Poisson Distribution

- Is Y ~ Poisson(λ*t*) skewed? Why or why not? For which values of Y?
- What does Y ~ Poisson(λ*t*) look like as λ increases?



Recall: Poisson Regression

- This leads us to the Poisson regression model

$$\log(Y) = \beta_0 + \beta_1 x_i + \ldots + \beta_p x_p + \log(t)$$

with offset log(t) to account for person-time of exposure, and

with linear function of covariates log(λ)

- Poisson regression then estimates the *Incidence Rate Ratio* comparing 2 populations for a 1-unit increase in x:

$$\textbf{IRR} = \frac{\lambda_1}{\lambda_2} = \frac{\mathrm{IR}_{\mathrm{population1}}}{\mathrm{IR}_{\mathrm{population2}}} = e^{\beta} = \log(Y)$$

- Model selection, assessment of confounding and effect modification, influence diagnostics, GOF, likelihood ratio testing, AIC and BIC—all are the same as in linear and logistic regression.

Special Example: Childhood Leukemia

- In the 1970s, 12 cases of childhood leukemia (age ≤ 19 years) occurred in Woburn, MA – there was concern about environmental toxins in the water supply

- Dr. Marvin Zelen (one of the founders of the Dept of Biostatistics) and Dr. Steve Lagakos (former Chair of the Dept) launched a study into this possible connection – later became a book, A Civil Action which was made into a movie!

- If the incidence rate of leukemia in the US in this age group is 5 cases per 10^5 person-years, is this an unusual occurrence?

- There were approximately 12,000 children age ≤ 19 years in Woburn over the 10-year period 1970-1979, resulting in t = 120,000 person-years of follow-up

Example: Childhood Leukemia

- The observed incidence rate of leukemia in Woburn was 12/120,000 = 10 cases per 10^5 person-years
- This above, is $\hat{\lambda}$, the point estimate for λ, where λ is the true incidence rate of leukemia in Woburn
- Let's first find a 95% confidence interval for λ
- Assume that the number of cases of leukemia in Woburn, represented by Y, follows a Poisson distribution

Example: Childhood Leukemia

- To find a confidence interval for λ, first consider a 95% CI for $\mu = \lambda t$, the *expected number of cases* in Woburn over a 10-year period

- Once we find the interval (μ_1, μ_2), we use the formula $\mu = \lambda t$ to get a 95% CI for λ, which is $(\mu_1/t, \mu_2/t)$, by solving for λ

- The point estimate for $\mu = \lambda t$ is the observed number of cases of leukemia, or $\hat{\mu} = 12$

CI for Expected Cases μ: Leukemia

- The 95% CI is (μ_1, μ_2), where μ_1 and μ_2 are the solutions to the equations:

$$P(Y \geq k \mid \mu = \mu_1) = \sum_{i=k}^{\infty} e^{-\mu 1} \mu_1^{i} / i! = 0.025 = \alpha/2$$

$$P(Y \leq k \mid \mu = \mu_2) = \sum_{i=0}^{k} e^{-\mu 2} \mu_2^{i} / i! = 0.025 = \alpha/2$$

- This is an *exact* Poisson confidence interval for μ

CI for Expected Cases μ: Leukemia

- To obtain the exact Poisson confidence interval for μ when we observe *x* events and the number of person-years is unknown (but large), we use:

- **`cii means 1 x, poisson`**

- **`cii means 1 12, poisson`**

```
-- Poisson  Exact --
    Variable |    Exposure          Mean     Std. Err.         [95% Conf. Interval]
-------------+------------------------------------------------------------------------
             |           1            12      3.464102         6.200575     20.96159
```

CI for Incidence Rate λ: Leukemia

- To obtain exact Poisson confidence limits for λ when there are *x* events in *y* person-years, we use:
- **`cii means y x, poisson`**
- **`cii means 120000 12, poisson`**

```
-- Poisson  Exact --
    Variable |    Exposure          Mean    Std. Err.        [95% Conf. Interval]
-------------+--------------------------------------------------------------------
             |      120000         .0001     .0000289         .0000517    .0001747
```

Poisson CI for Incidence Rate: Leukemia

- A 95% confidence interval for the expected number of events μ in Woburn is (6.2, 21.0)

- A 95% confidence interval for the incidence rate λ is (6.2/120,000, 21.0/120,000) or (5.2 / 10^5, 17.5 / 10^5) in rate of events per person-year

- Note that this interval does not contain 5.0/10^5, the incidence rate of childhood leukemia in the United States as a whole

- *This suggests that the number of cases observed in Woburn is unusually high*

A Civil Action

Two-Sample Example: Family History of Breast Cancer

- A sub-study was performed within the Nurses' Health Study looking at the association between family history of breast cancer and breast cancer incidence over a 15-year period

- Family history of breast cancer was determined at study enrollment

- Different women participated in the study for different lengths of time

14

Two-Sample Example: Family History of Breast Cancer

Family history of breast cancer	Number of new cases of breast cancer	Person-years of follow-up	Incidence rate (per 10^5 person-years)
Yes	295	80,539	366.28
No	1919	1,059,633	181.10

Two-Sample Example: Family History of Breast Cancer

- We wish to determine whether there is a significant difference in the incidence rates of breast cancer for women with and without a family history of breast cancer

- We also wish to estimate the magnitude of this effect using the *Incidence Rate Ratio*

$$\mathbf{IRR} = \frac{\lambda_1}{\lambda_2}$$

16

Two-Sample Hypothesis Test

- Suppose we have two independent groups of study subjects with t_1 and t_2 person-years of follow-up respectively, and corresponding incidence rates λ_1 and λ_2

- We wish to test the hypotheses:

$$H_0\colon\ \lambda_1 = \lambda_2$$

$$H_1\colon\ \lambda_1 \neq \lambda_2$$

17

Two-Sample Hypothesis Test

- We observe Y_1 events in group 1 and Y_2 events in group 2, from Poisson distributions with parameters λ_1 and λ_2, respectively

- If H_0 is true, we expect that $t_1 / (t_1 + t_2)$ of the total $Y_1 + Y_2$ events occur in group 1, and $t_2 / (t_1 + t_2)$ occur in group 2

- Note that $t_1 / (t_1 + t_2) + t_2 / (t_1 + t_2) = 1$

Two-Sample Hypothesis Test

- Therefore, the expected numbers of events in each of the two groups are:

$$E_1 = (Y_1 + Y_2) \times [t_1 / (t_1 + t_2)]$$

$$E_2 = (Y_1 + Y_2) \times [t_2 / (t_1 + t_2)]$$

- Conditional on the total number of events $n = Y_1 + Y_2$, the distribution of Y_1 is binomial with parameters n and $p_0 = t_1 / (t_1 + t_2)$ (see appendix)

Two-Sample Hypothesis Test

- Assuming a normal approximation to the binomial distribution (see appendix) is valid, the test statistic is:

$$Z = (Y_1 - np_0) / \sqrt{np_0q_0}$$

- The test statistic has a standard normal distribution if the null hypothesis is true

- To use the normal approximation, we should have that $np_0q_0 \geq 5$ (a "rule of thumb")

Example: Family History of Breast Cancer

- For these data,

$$n = 295 + 1919 = 2214.$$

$$p_0 = \frac{80,539}{80,539 + 1,059,633} = 0.071.$$

$$Y_1 = 295.$$

- If the null hypothesis is true, we have approximately

$$\begin{aligned} Y_1 &\sim N\left[2214(0.071),\ 2214(0.071)(0.929)\right] \\ &= \mathrm{N}(156.4,\ 145.3). \end{aligned}$$

Example: Family History of Breast Cancer

- The test statistic is then:

$$
\begin{aligned}
Z &= (Y_1 - np_0) \,/\, \sqrt{np_0q_0} \\
&= (295 - 156.4) \,/\, \sqrt{145.3} \\
&= 11.5
\end{aligned}
$$

- For the standard normal distribution, the probability of obtaining a value larger than 11.5 or smaller than –11.5 is $p < 0.001$

Example: Family History of Breast Cancer

- Thus, we reject the null hypothesis that the incidence rates are the same, and conclude that women with a family history of breast cancer have a significantly higher incidence of breast cancer

- If the normal approximation to the binomial distribution cannot be applied or we don't want to use an approximation, exact binomial tests do exist

Estimation of the Incidence Rate Ratio

- A point estimate of the incidence rate ratio IRR = λ_1 / λ_2 is given by the ratio of the two estimated incidence rates based on the samples of data

$$\widehat{IRR} = \hat{\lambda}_1 / \hat{\lambda}_2$$

- To obtain a confidence interval for IRR, we note that the logged IRR estimate is more closely normally distributed than the IRR estimate itself

- Therefore, we first find a CI for log(IRR) (see appendix)

Alternate Approach: Poisson Regression

- Here we have two Poisson-distributed outcomes, and a covariate (history of breast cancer in the family)

- Poisson regression methods (large sample) can also be used to estimate the IRR

- Note that the number of events Y_i does not follow a binomial distribution, because t_i varies for each subject and we don’t have a “denominator” (e.g., *X* successes out of *n* trials), so we cannot use logistic regression. The outcomes are not binary or binomial.

Poisson Regression

- As we have already seen, Poisson regression can be applied when the response variable is a count that follows a Poisson distribution (the analysis of incidence rates)

- It is not applicable for recurrent events where the probability of a second event is increased after a first event occurs (e.g., repeat hospitalizations for the same patient).

- However, it could be applicable for recurrent events where the incidence rate stays the same over time (e.g., number of migraine headaches over a particular time period).

Poisson Regression

- In some applications, the counts might only take on values 0 and 1 (e.g., when we are studying incidence rates of breast cancer, each subject either never gets the event during their follow-up period (Y = 0) or gets the event (Y = 1) but then follow-up stops after the event occurs), because it makes no sense to consider additional events occurring later with the same incidence rate.

- In some applications, counts and person time of follow-up is added up across subjects having the same set of covariates.

Poisson Regression Model

- Regarding the *offset:*

  - We've already seen that it quantifies the different follow-up times for each subject

  - When fitting Poisson regression models, the regression coefficient for the offset is fixed at 1 (there is no β coefficient for the offset term)

Poisson Regression

- The incidence rate can be a function of covariates, but should be constant over time, conditional on the values of those covariates (stationarity)

- We can control for multiple covariates, including both categorical and continuous covariates, as well as additive forms of potentially nonlinear covariates

- Follow-up times are allowed to vary for individual subjects

- Recall that the number of events for each subject is assumed to be independent of other subjects

Poisson Regression

- Suppose we have Poisson-distributed outcomes $Y_1, Y_2, \ldots, Y_n$ with a single covariate $x_1, x_2, \ldots, x_n$.

- If all observations have the same observation length, then the Poisson regression model assumes (with no time offset)

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

- If instead observations have possibly different observation lengths t_i, then the Poisson regression model assumes

$$\begin{aligned} E(Y_i) &= t_i \times \exp(\beta_0 + \beta_1 x_i) \\ &= \exp(\beta_0 + \beta_1 x_i + \log(t_i)) \end{aligned}$$

with offset $\log(t_i)$

Poisson Regression

- Here the *IRR* for a one unit increase in the covariate x_i is given by

$$\begin{aligned} \mathrm{IRR} &= \exp(\beta_0 + \beta_1 (x_i+1)) / \exp(\beta_0 + \beta_1 x_i) \\ &= \exp(\beta_1) \end{aligned}$$

$\rightarrow \log(\mathrm{IRR}) = \beta_1$

- As for logistic regression, maximum likelihood is used to estimate the β coefficients and to find standard error estimates, CIs, and p-values – all afforded by the GLM/Exponential Family machinery

Poisson Regression

- For the family history of breast cancer study, the first event has 295 cases, covariate = 1 (yes for family history), and offset = log(80,359).

- The second event has 1919 cases, covariate = 0 (no for family history), and offset = log(1059633).

- Using any computer package, we find →

Poisson Regression

Parameter	**95% CI**	**p-value**
$\hat{\beta}_0$ = -6.31	(-6.36, -6.27)	< 0.001
$\hat{\beta}_1$ = 0.707	(0.584, 0.829)	< 0.001

- For the incidence rate ratio we find $\widehat{IRR}$ = exp(0.707) = 2.03 with 95% CI (exp(0.584), exp(0.829)) = (1.79, 2.29).

- Poisson regression can be extended to allow more covariates and model building.

33

Example: Age at First Birth

- An analysis was performed relating age at first child birth (aafb) to breast cancer incidence among women in the Nurses' Health Study

- It was considered important to control for current age in the analysis, since breast cancer incidence increases with age and older women might have a different distribution of age at first birth than younger women

Example: Age at First Birth

- Follow-up for this study was over a 14 year period
- The subjects were followed for varying lengths of time from enrollment until the earliest of:

  (a) end of study

  (b) death

  (c) diagnosis of breast cancer

  (d) loss to follow-up

Example: Age at First Birth

Age	Nulliparous	Parous aafb ≤ 24	Parous aafb 25-29	Parous aafb ≥ 30
30-39	13 / 15,265* (85)	65 / 98,207 (66)	73 / 82,959 (88)	12 / 12,816 (94)
40-49	44 / 30,922 (142)	352 / 220,620 (160)	290 / 173,714 (167)	92 / 39,972 (230)
50-59	102 / 35,206 (290)	330 / 158,600 (208)	454 / 181,244 (250)	189 / 57,353 (330)
60-69	32 / 11,594 (276)	75 / 30,475 (246)	148 / 53,450 (277)	89 / 20,253 (439)

* Cases / person-years (incidence per 10^5 person-years)

aafb = age at first birth

Example: Age at First Birth

- We would like to compare breast cancer incidence:

(a) between parous (have had birth) and nulliparous (have had no birth) women, while controlling for current age

(b) by age at first birth among parous women, while controlling for age

Example: Age at First Birth

- We first fit the model with a single continuous covariate

$$\log(\lambda_i) = \alpha + \beta_1 x_{i1}$$

- $$\begin{aligned} x_{i1} = \text{ age } &= 35 \quad \text{if age 30-39} \\ &= 45 \quad \text{if age 40-49} \\ &= 55 \quad \text{if age 50-59} \\ &= 65 \quad \text{if age 60-69} \end{aligned}$$

- The incidence rate ratio associated with a 1 year increase in age is $\exp(\beta_1)$.

- Poisson regression uses cases as the outcome and log(person-years) as an offset.

Poisson Regression

IRR estimate	95% CI	P-value
1.041	(1.036, 1.046)	<0.001

- The IRR estimate is the exponentiated β estimate.
- The IRR estimate associated with a 10 year increase in age is $1.041^{10} = 1.50$ with 95% CI = $(1.036^{10}, 1.046^{10}) = (1.43, 1.57)$.
- Now look at the effect of being nulliparous (versus not nulliparous) on breast cancer incidence, controlling for age

Poisson Regression

IRR estimate	95% CI	P-value
1.041 for age	(1.036, 1.046)	<0.001
1.023 for nullip	(0.882, 1.187)	0.76

- Nulliparous status does not appear to be an independent predictor of breast cancer incidence, (nor does it appear to be a confounder of the effect of age on outcome (since crude IRR = 1.041 is very close to the adjusted IRR = 1.041 for age))

Example: Age at First Birth

- Is there effect modification occurring?

- That is, is the relationship between parity and breast cancer incidence different depending on a woman’s age? Or is the relationship between age and breast cancer incidence difference depending on a women’s nulliparous status? (Interaction works both ways)

- We can explore this possibility by creating an interaction term, the product of age × nulliparity

Poisson Regression

IRR estimate	95% CI	P-value
1.041 for age	(1.036, 1.046)	<0.001
0.923 for nullip	(0.373, 2.286)	0.86
1.002 for interact	(0.985, 1.019)	0.82

- Nulliparous status is not an effect modifier of the effect of age on breast cancer incidence, nor is age an effect modifier of the effect of parity on outcome (since the P-value for the interaction term is 0.82)

Example: Age at First Birth

- From previous literature it is known that breast cancer incidence rises more steeply before menopause (around age 50), than after menopause

- Therefore, it might be preferable to express age as a categorical variable rather than a continuous variable

- Since there are 4 age categories, 3 indicator variables are created

Poisson Regression

IRR estimate	95% CI	P-value
2.147 for age45	(1.814, 2.542)	<0.001
3.191 for age55	(2.706, 3.762)	<0.001
3.811 for age65	(3.163, 4.592)	<0.001
1.032 for nullip	(0.890, 1.197)	0.68

- Baseline category: age 30-39

Example: Age at First Birth

- Compared with the reference group (age group 30-39), the estimated incidence rate ratios are 2.1, 3.2, and 3.8 for women in age groups 40-49, 50-59, and 60-69 respectively
- The steepest rise in incidence is from 30-39 years to 40-49, with somewhat smaller increases after age 50
- All three of the older age groups have incidence rates that are significantly higher than the youngest age group (all $p < 0.001$)

Example: Age at First Birth

- In addition, nulliparous women still have a higher incidence of breast cancer than parous women, estimated IRR = 1.03, comparable to the IRR in the model with continuous age, but is again not statistically significant (*p* = 0.68)

Example: Age at First Birth

- It is also known from previous studies that the incidence of breast cancer for parous women varies with age at first birth (aafb)

- Therefore, we use the model:

$$\ln(\lambda) = \alpha + \beta_1 age40-49 + \beta_2 age50-59 + \beta_3 age60-69$$
$$+ \beta_4 aafb <= 24 + \beta_5 aafb25-29 + \beta_6 aafb30+$$

- We have omitted the nulliparous term since it is the reference group (aafb = 0 for nulliparous women)

Poisson Regression

IRR estimate	95% CI	P-value
2.128 for age45	(1.798, 2.520)	<0.001
3.067 for age55	(2.600, 3.618)	<0.001
3.563 for age65	(2.954, 4.298)	<0.001
0.853 for aafb2	(0.728, 0.999)	0.048
0.971 for aafb3	(0.831, 1.134)	0.707
1.332 for aafb4	(1.119, 1.585)	0.001

- Baseline categories: age 30-39, nulliparous status

Example: Age at First Birth

- After controlling for age, the IRR is 0.85, 95% confidence interval (0.73, 1.00), *p* = 0.048 for women with age at first birth ≤ 24 versus nulliparous women

- The comparable IRRs for women with aafb 25-29 and 30+ are 0.97 (*p* = 0.71) and 1.33 (*p* = 0.001) versus nulliparous women

- Hence, women with an early aafb are at decreased risk of breast cancer, and women with a late aafb are at increased risk versus nulliparous women

Poisson Regression

β estimate	95% CI	P-value
0.755 for age45	(0.586, 0.924)	<0.001
1.121 for age55	(0.956, 1.286)	<0.001
1.271 for age65	(1.083, 1.458)	<0.001
-0.159 for aafb2	(-0.317, -0.001)	0.048
-0.030 for aafb3	(-0.185, 0.126)	0.707
0.287 for aafb4	(0.113, 0.460)	0.001
-7.095 for const	(-7.301, -6.889)	<0.001

Prediction from Poisson Regression Models

- Suppose we want to estimate the 5-year incidence of breast cancer for 45-year-old women with age at first birth 22 years

- From the specification of the model, we have:

$$\ln(\hat{\mu}) = \hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_4 + \ln(5)$$

$$= -7.095 + 0.755 - 0.159 + 1.609 = -4.890.$$

$$\text{Thus, } \hat{\mu} = \exp(-4.890) = 7.52 \text{ x } 10^{-3} = 752/10^5.$$

Next: GOF in Poisson Regression

- Recall the use of ‘calibration’ in assessing GOF when regression modeling

- Suppose we wish to test the hypothesis (sounds familiar!):

H_0: the model fits the data (well-calibrated) versus

H_A: the model does not fit the data

- J = number of covariate patterns
- O_j = observed number of events for the j^{th} covariate pattern, $j = 1,\ldots, J$
- E_j = expected number of events under the Poisson model for the j^{th} covariate pattern, $j = 1,\ldots, J$

Poisson Goodness-of-Fit

- There are two summary statistics used to test goodness-of-fit (calibration) of Poisson regression models:

  - Pearson chi square statistic

  - Deviance statistic

- Both of these are useful only when you have a small to moderate number of covariate patterns *J* (there is not an assessment for a large number of covariate patterns)

Pearson Chi Square Statistic

- The Pearson chi square statistic is similar to the chi-square goodness-of-fit test statistic for logistic regression, and is given by

$$X^2_{Pearson} = \sum_{j=1}^{J} \frac{\left(O_j - E_j\right)^2}{E_j} \sim \chi^2_{J-K} \ under\ H_0,$$

$$\text{p-value} = \Pr(\chi^2_{J-K} > X^2_{Pearson}).$$

where *K* is the number of parameters estimated from the data (including the intercept), and *J* is the number of covariate patterns

- It asks ‘How off were our predictions by assuming the Poisson model holds?’

- As before, small p-values correspond with lack of fit

Pearson Chi Square Statistic

- For the breast cancer data, there are 16 covariate patterns (4 age groups × 4 age at first birth groups) and *k* = 7 parameters estimated from the data (constant term, 3 age indicators, and 3 aafb indicators)

- Therefore, the test statistic has 16 – 7 = 9 df

$$X^2_{Pearson} \sim \chi^2_9 \text{ under } H_0.$$

Pearson Chi Square Statistic

```
. estat gof

          Pearson goodness-of-fit  =  8.686447
           Prob > chi2(9)           =    0.4667
```

- Since p = 0.47, we are unable to reject the null hypothesis

- This indicates adequate fit of the model with age categories and aafb categories

- What about the model with continuous age?

Pearson Chi Square Statistic

- Based on a model using linear age plus three indicator variables for aafb status (5 parameter model, so 16 – 5 = 11 d.f.):

```
. estat gof

      Pearson goodness-of-fit  =  41.49026
      Prob > chi2(11)          =    0.0000
```

- Here $p < 0.0001$, and we reject the null hypothesis
- This indicates that the model with continuous age is not well calibrated to the observed data – keeping categorical age seems more reasonable

Deviance Goodness-of-Fit Statistic

- Next, the deviance goodness-of-fit statistic is given by:

$$X^2_{Deviance} = 2\sum_{j=1}^{J} O_j \ln(\frac{O_j}{E_j}) \sim \chi^2_{J-K} \ under \ H_0 .$$

- It has the same degrees of freedom as the Pearson test, and is equivalent to the likelihood ratio test comparing your model with the saturated model (containing *J* parameters)

- It asks ‘How do our predicted counts match up with the true counts, or saturated model?’

Deviance Statistic

Categorical age:

```
. estat gof

      Deviance goodness-of-fit =  8.716593
      Prob > chi2(9)           =    0.4638
```

Continuous age:

```
. estat gof

     Deviance goodness-of-fit =   43.5129
     Prob > chi2(11)          =    0.0000
```

So we get basically the same results as for the Pearson chi square goodness-of-fit test

Deviance Statistic via LRT

- Equivalent to the deviance statistic, one could perform a likelihood ratio test comparing your model to the saturated model and you will get an equivalent result (same test statistics, same degrees of freedom)

- Usually the Pearson and deviance statistics will generally give you similar findings, though they *can only be used for small to moderate number of covariate patterns*

Model Building

- Model building and model interpretation with Poisson regression follows procedures analogous to that of logistic regression – selection procedures, interpretation of IRR's (rather than OR's), confounding and effect modification, likelihood ratio tests comparing nested models, AIC and BIC values, etc.

When can the Poisson Model Break Down?

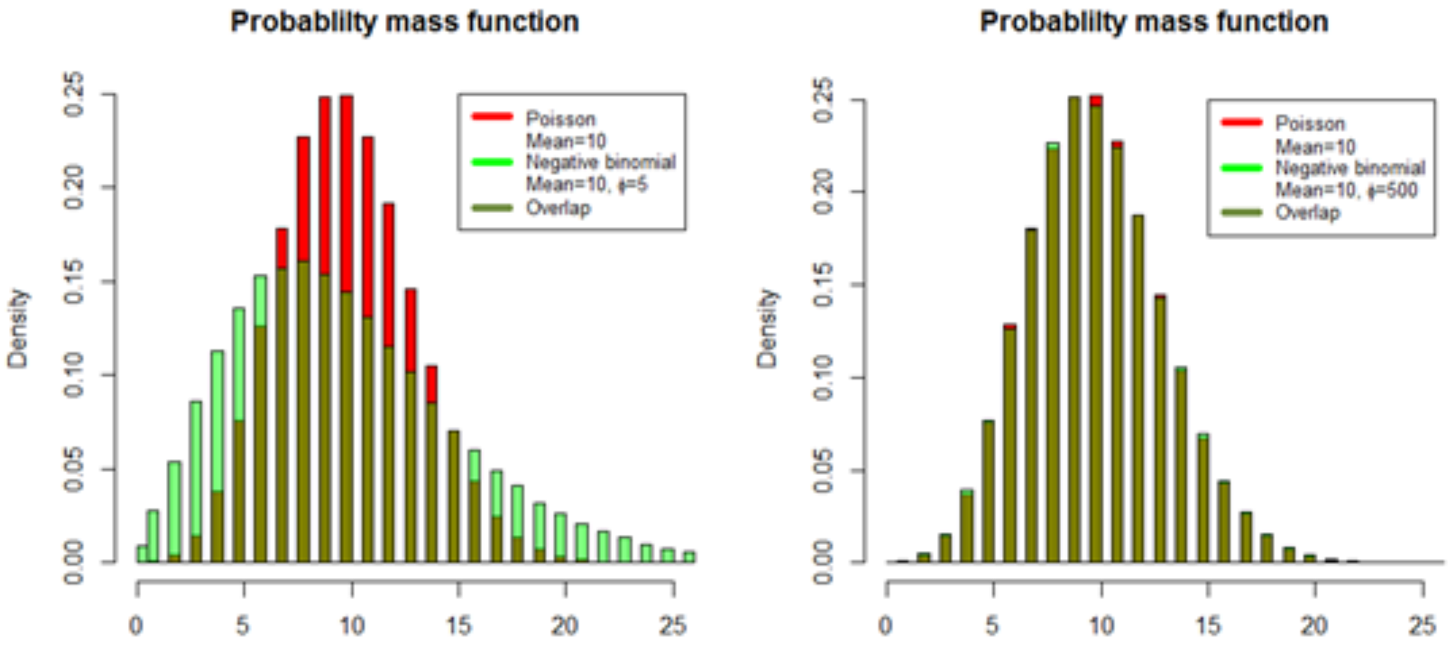
- (1) In cases of *overdispersion*, when there is more variability in the data than equals the mean (as is specified by the Poisson model) – can be caused by dependence between counts across individuals/covariate patterns, or that the IR varies over time.

  - underestimating variability leads to invalid inference

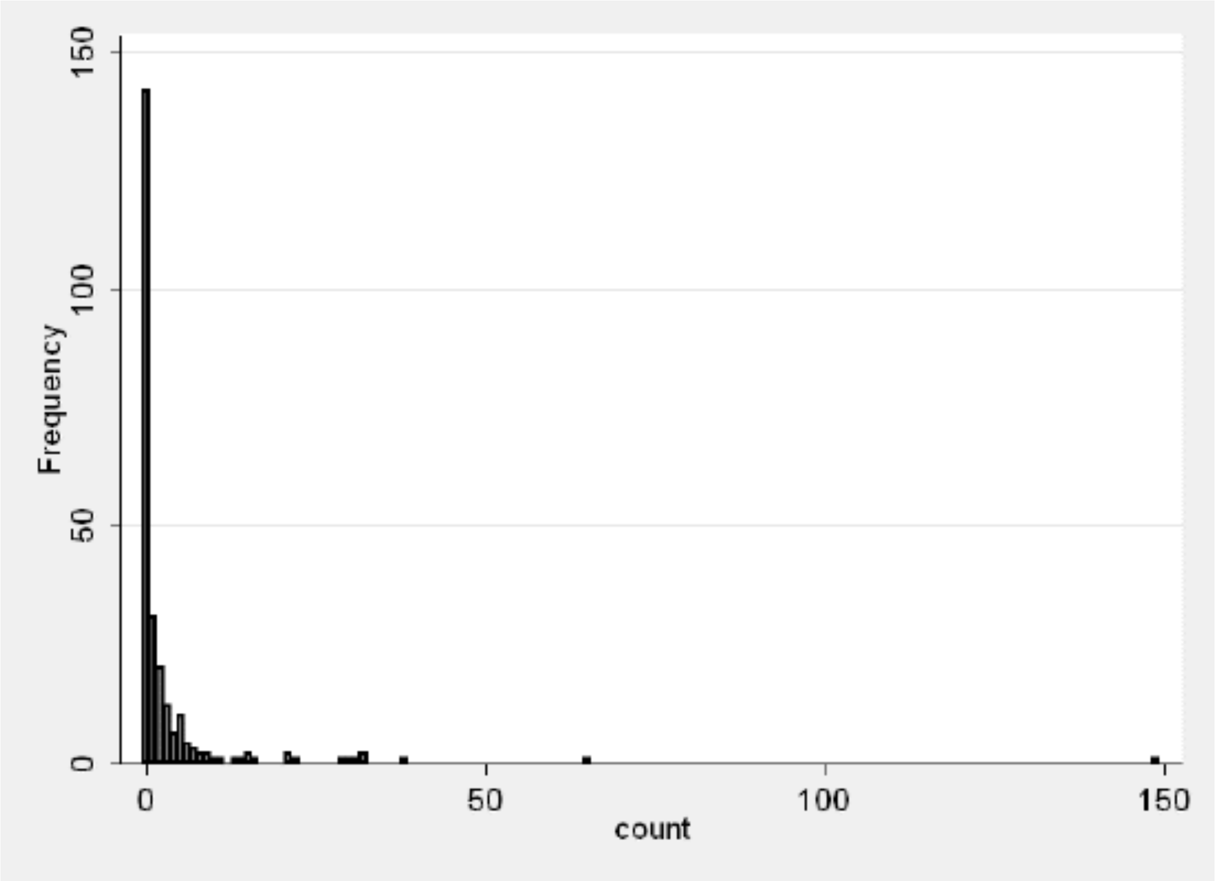
- (2) When there are a lot of zeros (or a lack of zeros); more (less) than expected for the Poisson model

- What does this all look like?

(1) When can the Poisson Model Break Down?



(2) When can the Poisson Model Break Down?



Need for Extensions

That is,

- One (strong) assumption of Poisson regression is that the mean and variance of a Poisson outcome are equal. Sometimes this assumption is violated (often with higher variances than means), and you might need a *negative binomial* model or an *overdispersed* Poisson regression model

- In other cases, you might have more “zero outcomes” than expected, and might need to fit a *zero-inflated* Poisson model.

Need for Extensions

What to do?

(1) Check for overdispersion using GOF model

- (Pearson or Deviance)
- in lab you will work through how to do this even with large number of covariate patterns by rule of thumb approach

(2) Employ a Negative Binomial or Zero-Inflated Poisson model, possibly with robust variance estimation

Zero-Inflated Poisson Regression

Here is one solution!

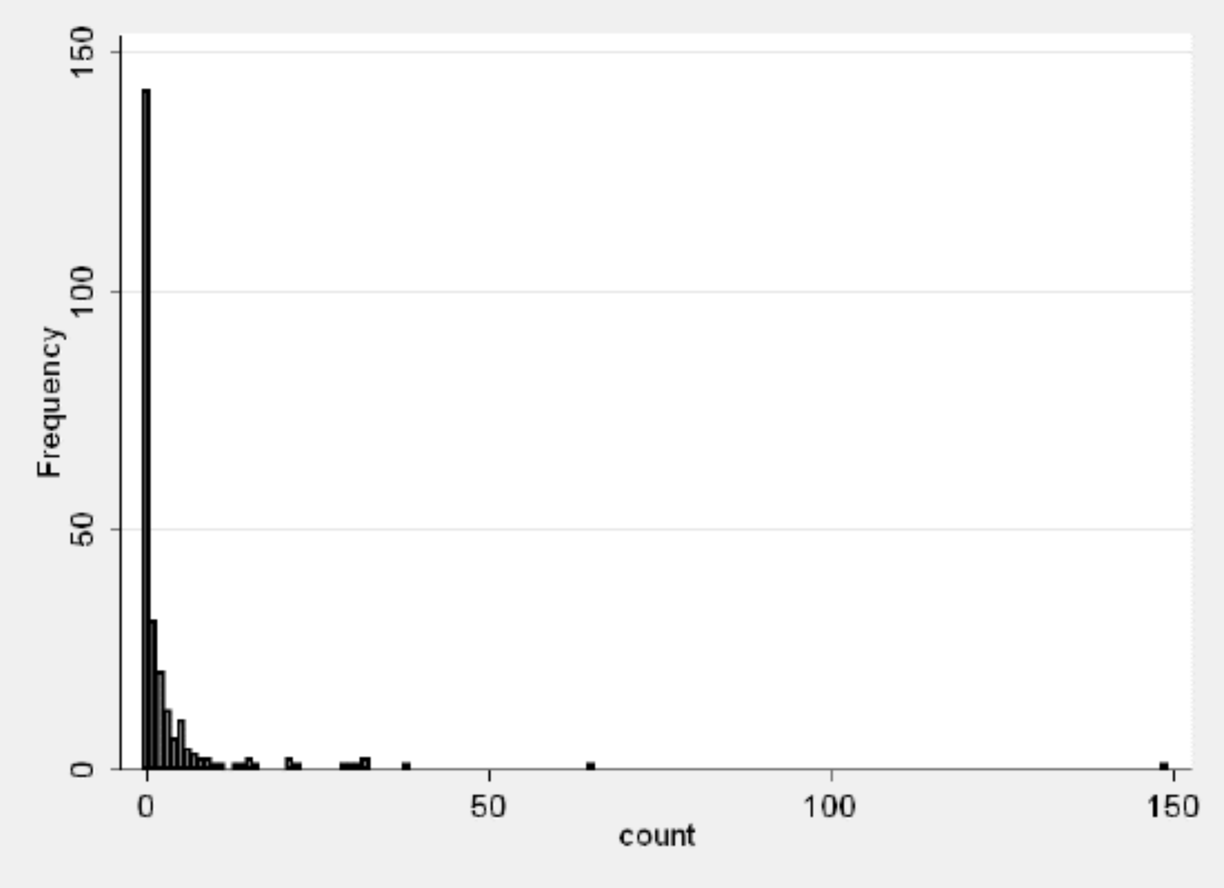
- The Zero-Inflated Poisson regression model (ZIP) allows excess zero counts, i.e., more zeros than expected under the usual Poisson model (similarly the Zero-Truncated Poisson regression model allows for lack of zeros, but is less commonly used)

- Two components of the process are modeled.

  - (i) First, there is a binary distribution that generates some ‘structural zeros’.

  - (ii) Second, there is a Poisson distribution that generates counts, some of which may also be zero, since Poisson outcomes can be 0, 1, 2, …

Zero-Inflated Poisson Regression



Zero-Inflated Poisson Regression

- In the simplest case with no covariates, let p = probability of a structural zero (binary outcome), and λ = Poisson incidence rate. Then we have:

(binary piece): $P(Y = 0) = p + (1 – p) \exp(– \lambda)$

(Poisson piece): $P(Y = i) = (1 – p) \lambda^{i} \exp(– \lambda) / i!$ $\quad i = 1, 2, 3, \ldots$

- Here

$$E(Y) = p\,(0) + (1 – p)\,\lambda = \lambda\,(1 – p)$$

$$E(Y^2) = (1 – p)\,(\lambda + \lambda^2)$$

$$\text{Var}(Y) = E(Y^2) – [E(Y)]^2 = \lambda\,(1 – p)\,(1 + \lambda p)$$

$$> E(Y) \text{ when } p > 0$$

Zero-Inflated Poisson Regression

- Next, we could extend this model to include covariates in two ways: First, we could use logistic regression (with covariates) to model *p*. Second, we could use Poisson regression (with covariates and possibly different exposure times) to predict λ.

- The covariates used for each of the models do not have to be the same. Maximum likelihood methods can be used to fit the parameters.

Zero-Inflated Poisson Regression

- Using different exposure times for different covariate patterns or summing Poisson counts across subjects (like in the example we just did) could likely make the use of a ZIP less appropriate.

- Using ZIP only makes sense if this idea about having structural zeroes for some outcomes is reasonable.

- The ZIP model does not fall in the generalized linear model family. ☹

Zero-Inflated Poisson Regression

- It can happen that the zip likelihood being maximized in not concave, especially when the same covariate is used in both parts of the model, and so convergence of the modeling is not guaranteed

- Because of this, many people who use zip select different covariates to predict the structural zeroes and the Poisson rates

- An extension (in a different direction!) is the use of zero-truncated Poisson regression (where we never get to observe a zero outcome)

Negative Binomial Regression

- Going back to Poisson regression, we are modeling

  $P(Y_i = k) = \exp(-\mu_i)\,\mu_i^k / k!$ for $k = 0, 1, 2, \ldots$

where $\mu_i = \exp(\beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + \log(t_i))$

and $E(Y_i) = \mu_i$ and $\mathrm{Var}(Y_i) = \mu_i$.

- Here, the mean and variance of all observations are the same, as expected for the Poisson distribution.

Negative Binomial Regression

- Negative binomial regression can be motivated by a gamma mixture of Poisson probabilities to develop a model with extra-Poisson variation (details omitted). With this model, we have

$$E(Y_i) = \mu_i = \exp(\beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + \log(t_i))$$

and

$$\mathrm{Var}(Y_i) = \mu_i \times (1 + \alpha\, \mu_i) > E(Y_i) \text{ when } \alpha > 0$$

- (Here α now relates to the overdispersion, not the intercept term)

Negative Binomial Regression

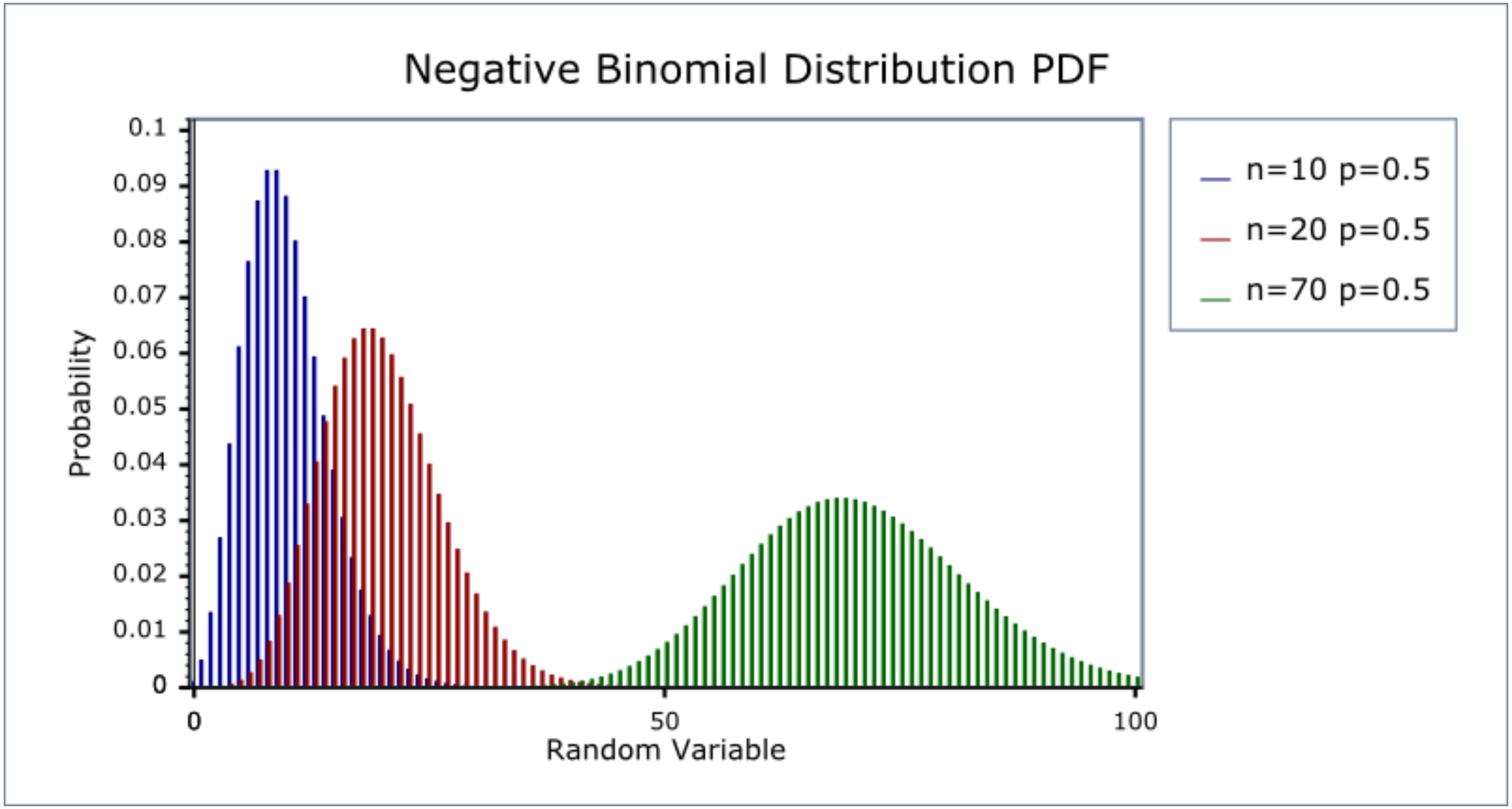
- This model can be estimated via maximum likelihood

- Exponentiated β coefficients still have IRR interpretations

- We can formally test whether H_0: $\alpha = 0$ or not to assess overdispersion

- The negative binomial regression model is included within the generalized linear model family ☺

Negative Binomial Regression



Robust Variance Estimation

- When fitting binary outcomes with logistic regression, we develop a model for the mean of the binary outcomes (E(Y_i) = p_i). Since the outcomes are binary, for the true model we know that the variances of the binary outcomes are $p_i\,(1 - p_i)$.

- When fitting count outcomes with Poisson regression, we are basically assuming that E(Y_i) = Var(Y_i) when we are getting our ML estimates. In practice, you might have the correct model for the mean, but the variances could be different (often larger) than that anticipated by the Poisson model.

Robust Variance Estimation

- Independently, Huber and White developed empirical variance estimators (for the point estimates) that are appropriate even under overdispersion (or underdispersion), provided the mean model is correctly specified. These are sometimes called robust variances, or sandwich variances (because of the form of the calculation).

- It often makes sense to compare the robust variance estimates to the usual (MLE information-based) variance estimates. (Or possibly, to consider whether zip or negative binomial modeling is appropriate.)

Robust Variance Estimation

- These robust variance estimates can be extended to multivariate or longitudinal responses, where we call this the “generalized estimating equations” or GEE approach.

- Basically, we calculate the parameter estimates assuming working independence, and calculate robust (or GEE) variance estimates, and use these to construct confidence intervals or *p*-values. (This can be extended to other working assumptions.)

Coming Up

- Survival analysis!

**** Appendix ****

- Distributional approximations:
  - Poisson approximation to the Binomial
  - Normal approximation to the Poisson

- Confidence intervals for the incidence rate (large sample, and exact)

- Examples

- Two group comparisons and incidence rate ratios

Poisson Approximation to Binomial Distribution

- Next we delineate several well known distributional approximations. Approximations tended to come about in the absence of the computing speed and capacities we have today, when most relied upon Z tables and such to 'look up' computation results.

- If $n \geq 100$ and $p < 0.01$ (this is a "rule of thumb"), we can approximate a binomial distribution with parameters n and p by a Poisson distribution with parameter $\mu = np$

- This works because the variance of a binomial random variable npq (ie np(1-p)) is approximately equal to the mean np if p is small and q is close to 1 (where $q = 1 – p$)

Poisson Approximation to Binomial Distribution

- Suppose that a hospital serves a population of exactly n = 25,000 people
- And the probability that any one of them is diagnosed with a certain rare cancer over the next year is $p = 3.6 \times 10^{-5}$
- Assuming a binomial distribution, $E(Y) = np = 0.9$
- To calculate the probability of observing at least 5 cases of cancer in 1 year, we could instead use the Poisson distribution with $\mu = np = 0.9$

Poisson Approximation to Binomial Distribution

- For X ~ Binomial(n = 25,000, p = 3.6 x 10^{-5}) then

  $P(X \geq 5) = 0.0023435$

- For X ~ Poisson(μ = 0.9) then

  $P(X \geq 5) = 0.0023441$

- These probabilities are very similar, and both are less than 0.05; therefore, observing 5 events in 1 year is unusual

- Note: Given computers, this approximation may not really be needed, but could be helpful in some situations

Normal Approximation to the Poisson

- Suppose there is a rare form of influenza which yields 15 cases per year on average in a certain community

- In 2017 there are 25 cases

- Is this an unusual occurrence?

- We could use the Poisson distribution to answer this question

- Y is a random variable representing the number of events in 1 year

Normal Approximation to the Poisson

- Note that $\lambda = 15$ cases per year and $T = 1$ year

- Therefore, $\mu = 15$

- We want to compute

$$P(Y \geq 25 \mid \mu = 15) = 1 - P(Y \leq 24)$$

$$= 1 - \sum_{k=0}^{24} e^{-15} (15)^{k} / k!$$

- We can also use a normal approximation to the Poisson distribution

Normal Approximation to the Poisson

- A Poisson random variable Y with parameters λ and T has mean $\mu = \lambda T$ and variance $\sigma^2 = \lambda T = \mu$

- We can approximate the Poisson Y by a normal random variable W with mean = variance = μ

- We can approximate $P(Y = k)$ by

  $P(k - 0.5 \leq W \leq k + 0.5)$ using a *continuity correction*

- This approximation should be used only if $\mu \geq 10$ (a “rule of thumb”)

Example: Influenza

- We approximate a Poisson random variable Y with parameter $\mu = 15$ by a normal random variable W with mean 15 and variance 15

- Thus P($Y \geq$ 25) is approximated by

$$
\begin{aligned}
P(W \geq 24.5) \quad &= P(Z \geq [24.5 - 15] / \sqrt{15}) \\
&= P(Z \geq 2.45) \\
&= 0.00714281
\end{aligned}
$$

- It is unusual to observe 25 cases in one year; there appears to be a significant increase in the number of cases in 2017

Example: Influenza

- Compare this with the exact Poisson calculations when Y ~ Poisson(15). Here, P(Y ≥ 25) = 0.01116478.

- This approximation is accurate to the second decimal place, and would be more accurate with larger μ (here μ is only 15)

- Note: Given computers, this approximation may not really be needed, but could be helpful in some situations

CI for Expected Cases (large sample)

- For a Poisson distribution, the mean and variance are both equal to μ
- If μ is sufficiently large, the Poisson distribution can be approximated by a normal distribution with mean and variance both equal to μ
- The point estimate for μ is the observed number of cases *x*

CI for Expected Cases (large sample)

- An approximate 95% confidence interval for μ is

$$x \pm 1.96\sqrt{x}$$

- Thus, if $x = 100$, the 95% confidence interval for μ would be

$$100 \pm 1.96\ (10) \quad \text{or} \quad (80.4,\ 119.6)$$

CI for Expected Cases μ

This approximate 95% confidence interval is similar to the Poisson exact interval of (81.4, 121.6)

The larger the value of μ, the closer the two intervals should be

```
cii means 1 100, poisson
```

```
                                                   -- Poisson  Exact --
    Variable |   Exposure    Mean    Std. Err.      [95% Conf. Interval]
-------------+---------------------------------------------------------
             |          1     100          10        81.36399   121.6268
```

Interval Estimation of the IRR

We have that:

$$\ln(IR\hat{R}) = \ln(\hat{\lambda}_1) - \ln(\hat{\lambda}_2)$$

Thus,

$$\text{var}[\ln(IR\hat{R})] \;\; = \text{var}[\ln(\hat{\lambda}_1)] + \text{var}[\ln(\hat{\lambda}_2)]$$

From the delta method,

$$\text{var}[\ln(\hat{\lambda}_1)] = 1/Y_1,\;\; \text{var}[\ln(\hat{\lambda}_2)] = 1/Y_2,$$

and

$$\text{var}[\ln(IR\hat{R})] = 1/Y_1 + 1/Y_2.$$

93

Interval Estimation of the IRR

- Therefore, a 95% CI for log(IRR) is

$$\log(\hat{\lambda}_1 / \hat{\lambda}_2) \pm 1.96 \sqrt{1/Y_1 + 1/Y_2}$$

- To get a 95% CI for the IRR itself, we exponentiate the upper and lower bounds of this interval

- If the incidence rates for the two groups are the same, we would expect the confidence interval to contain the value 1

Example: Family History of Breast Cancer

$$IR\hat{R} = (366/10^5)/(181/10^5) = 2.0.$$

A 95% CI for ln(IRR) is given by:

$$\ln(2.0) \pm 1.96\sqrt{1/295 + 1/1919}$$

$$= 0.704 \pm 1.96(0.063)$$

$$= 0.704 \pm 0.123 = (0.581,\ 0.827) \ = \ (\mathrm{c}_1, c_2).$$

The corresponding 95% CI for IRR is:

$$[\exp(0.581), \exp(0.827)] = (1.8, 2.3).$$

Example: Family History of Breast Cancer

- The 95% CI for the IRR does not contain 1, which demonstrates that the incidence rates for women with and without a family history of breast cancer are not the same

- The interval lies entirely above 1

- This interval should only be used if

$$np_0q_0 = (Y_1 + Y_2)\ \mathrm{T}_1\,\mathrm{T}_2 / (\mathrm{T}_1 + \mathrm{T}_2)^2 \geq 5$$

96