PM 592
Regression Analysis for Public Health Data Science

Week 11

Poisson & Negative Binomial Models

1

Poisson & Negative Binomial Models

Intro to GLM

Poisson Regression

Poisson Regression: GOF

Negative Binomial Regression

Rate Outcomes

2

Lecture Objectives

- Explain the concept of the link function and what types of transformations can be made
- Determine whether data is suitable for Poisson regression
- Interpret Poisson regression output
- Determine if an outcome is overdispersed
- Evaluate the fit of a Poisson regression model
- Explain when to use negative binomial regression
- Implement a Poisson or negative binomial model with a rate outcome

3

1. Review

✓ How to build a predictive model

✓ Ways of evaluating the diagnostic/prognostic ability of the model

✓ Determining the best cut point

✓ Explaining the ROC curve and AUROC

4

2. Introduction to GLM

We have already seen two types of linear models in this course:

1. Ordinary least squares regression
2. Logistic regression

The regression framework is desirable when modeling effects because:

- We can model the effects of **several independent variables** simultaneously
- We can control for **confounding** and examine **interaction** terms
- We have **flexibility** (linear, categorical, polynomial, etc.) in modeling variables
- We obtain **parameter estimates** with confidence intervals and significance values
- We can determine the **predicted** (expected) values of the outcome

5

2. Introduction to GLM

We can use linear models for several types of outcomes:

Continuous

- OLS regression
- ANOVA, ANCOVA

ANOVA and ANCOVA are special cases of OLS regression.

Binary

- Logistic regression
- Probit regression

Discrete/Count

- Poisson regression
- Negative binomial regression

6

Linear models contain a **random component**, a **systematic component**, and a **link function**:

$$g(E(Y)) = g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$
$$Y \sim ?$$

7

For OLS regression:

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

The random component $Y \sim N(\mu, \sigma^2)$

The link function $g(E(Y)) = E(Y) = \hat{Y}$

8

For Logistic regression:

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

The random component $Y \sim B(n, \pi)$

The link function $g(E(Y)) = logit(E(Y)) = logit(\hat{\pi})$

9

This allows us to use "linear" regression in a generalized way:

- A GLM doesn't assume that the raw X and Y are necessarily linearly related
- However, it is assumed that there is a linear relationship between the predictors and the transformed response (e.g., between ***X*** and $logit(\pi)$).

10

Let's look at something we're already familiar with.

Linear regression

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

We don't need to perform any transformation on this outcome.

We can use the predicted Y values as-is.

So, we use a very simple link.

The **identity link** is g(μ) = μ.

11

Here's how we make it "generalized"

All GLMs have the same systematic component, but they can have different link functions depending on the random component.

Regression Type	Statistical Model	Random Component	Link function
Linear	$\mu_i = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k$	Normal	Identity
Poisson	$\ln(\mu_i) = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k$	Poisson	Natural log
Logistic	$ln\left[\frac{\mu_i}{1-\mu_i}\right] = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k$	Binomial	Logit
Probit	$\phi^{-1}(\mu_i) = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k$	Binomial	(Normal CDF)$^{-1}$

12

2. Introduction to GLM 13

So why don't we just transform the Y variable, use linear regression as usual, and back-transform?

Here are a couple of reasons:

- Restrictiveness—The GLM approach doesn't assume normality or homogeneity of residual variance.
- Interpretability—If you transform Y, you have to interpret the regression coefficient on the transformed outcome variable.
- Accuracy—Simple transformations like this often don't achieve normality.
- Elegance—We are able to use the information about the distribution of Y in our modeling approach.

13

2. Introduction to GLM 14

Let's review: **The logit** is one function that will yield predictions constrained between 0 and 1.

$$\text{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$$

In extreme cases:

Y = 0: $\text{logit}(0) = \ln\left(\frac{0}{1-0}\right) = -\infty$

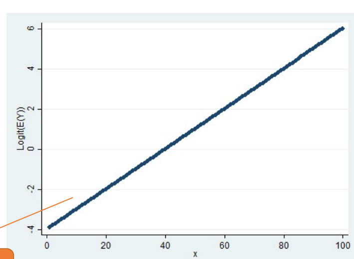
Y = 1: $\text{logit}(1) = \ln\left(\frac{1}{1-1}\right) = +\infty$

14

2. Introduction to GLM 15

Let's use the logit link:

$$\text{logit}(\pi) = \ln\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$



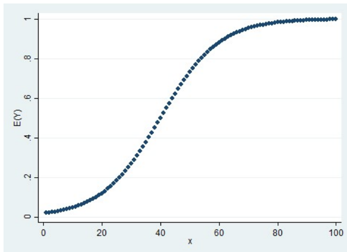
The prediction is "linear in the logit."

15

2. Introduction to GLM

What happens when the equation is transformed in terms of Y?

$$\pi = \frac{e^{\beta_0+\beta_1 x_1+..+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+..+\beta_k x_k}}$$



16

2. Introduction to GLM

Probit regression also yields predictions constrained between 0 and 1.

- Probit regression uses information about the normal probability density function.
- We know that the probability in this function is constrained between 0 and 1.
- ϕ^{-1} denotes the inverse cumulative density function.

Given a value z, what is the probability that Z<z?

While appropriate for modeling dichotomous outcomes, probit regression isn't as commonly used as logistic regression and we won't cover it in this course.

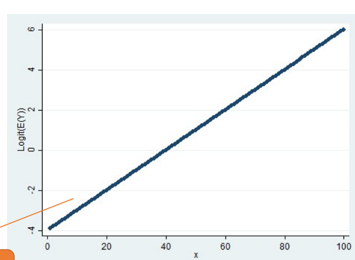
Probability unit = "probit"

17

2. Introduction to GLM

Let's use the probit link:

$$\text{probit}(\pi) =$$
$$\phi^{-1}(\pi) =$$
$$\beta_0 + \beta_1 x_1 + .. + \beta_k x_k$$



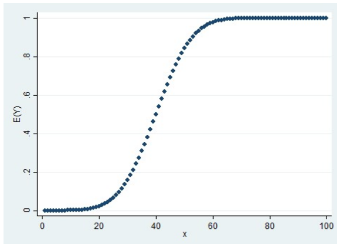
The prediction is "linear in the probit"

18

2. Introduction to GLM 19

What happens when the equation is transformed in terms of Y?

$$\pi = \phi(\beta_0 + \beta_1 x_1 + .. + \beta_k x_k)$$



19

2. Introduction to GLM 20

Recap

- Generalized linear models allow you to use a familiar regression modeling approach with outcomes that may have different distributions or a nonlinear relationship with the outcome.

20

2. Introduction to GLM 21

Recap

- Explain the three components of a generalized linear model

21

3. Poisson Regression
22

Poisson random variables (used for count data) contain data that are only positive.

Examples of count data:

- Number of publications produced by PhD students at different institutions
- Number of spam phone calls you receive on your cell phone each day
- Number of days an individual stays in recovery in the hospital

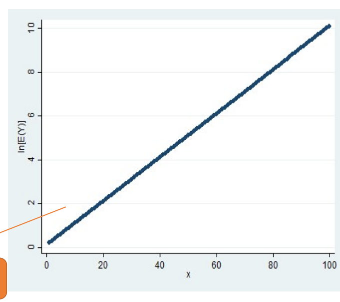
What type of link function (–∞, +∞) would map to predictions that are constrained between [0, +∞)?

22

3. Poisson Regression
23

Let's use a natural log link:

$$\ln(\mu) =$$
$$\ln\big(E(Y)\big) =$$
$$\beta_0 + \beta_1 x_1 + .. + \beta_k x_k$$



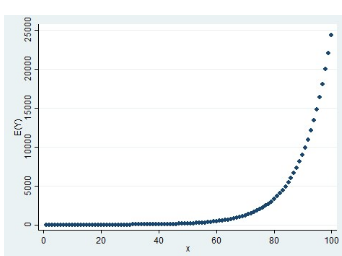
The prediction is "linear in the log."

23

3. Poisson Regression
24

What happens when the equation is transformed in terms of Y?

$$\mu =$$
$$E(Y) =$$
$$e^{\beta_0 + \beta_1 x_1 + .. + \beta_k x_k}$$



24

What distributions of Y are suitable for this type of analysis?

- Count data really comes from binomial processes (the probability of a certain number of successes, given some probability).
- The **Poisson limit theorem** states that the $\mathrm{Poisson}(\lambda)$ distribution is the limit of the $\mathrm{Binomial}(n, \pi)$ distribution with $\lambda = n\pi$ as $n \to \infty$.
- In this situation, λ is the expected number of events.

25

Example

An ER department performs 500 surgeries per month (n=500). On average, one surgery of the 500 will result in patient death (p=1/500). How can we model this?

1) X ~ Binomial(500, 1/500)
2) X ~ Poisson(1)

Notice, the λ parameter of the Poisson distribution is the **expected value**. That is, we would expect to observe 1 fatality in this department.

26

Example

There are 50,000 e-mails sent at USC KSOM in any given day (n=50,000). On average, 50 of these get sent to the wrong recipient due to user error (1/1000). How can we model this?

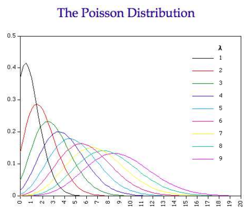
1) X ~ Binomial(50,000, 1/1000)
2) X ~ Poisson(50)

27

3. Poisson Regression 28

Under the Poisson distribution,

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



Again, this distribution has just one parameter:

$$\lambda = E(Y) = V(Y) > 0$$

This implies:

- The mean of a Poisson variable equals its variance.
- We expect more variation in Y when E(Y) is larger.
- For many count variables, λ is small and there are many observed zeroes.

As λ increases, the Poisson distribution approaches a normal distribution.

28

3. Poisson Regression 29

Back to our ER example:

500 surgeries per month, p=1/500 of patient death

What is the probability of observing no deaths in a given month?

$$P(Y = 0) = \frac{e^{-1}1^0}{0!} = 0.3679$$

What is the probability of observing 3 deaths in a given month?

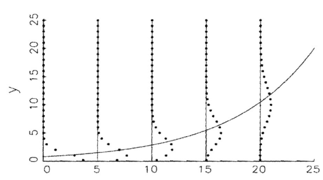
$$P(Y = 3) = \frac{e^{-1}1^3}{3!} = 0.0613$$

29

3. Poisson Regression 30

Some properties of **Poisson regression**

- $-\infty < \ln \mu_{Y|X} < \infty$ and $Y \sim \text{Poisson}(\mu)$
- This lets us use our linear predictor (with a range of -∞, ∞) to predict outcomes ranging from 0, ∞.
- We assume at each X, Y has a specific Poisson distribution with mean and variance as a function of X.



30

3. Poisson Regression 31

If $\ln \mu_{Y|X} = \beta_0 + \beta_1 x$ then we can solve for outcome directly:

$$\mu_{Y|X} = \exp(\beta_0 + \beta_1 x_1) = \exp(\beta_0) * \exp(\beta_1 x)$$

- When X=0, our expected count outcome is $\exp(\beta_0)$.
- A one-unit increase in X has a **multiplicative effect** on outcome, multiplying the baseline mean count by $\exp(\beta_1 x)$.
  - If β_1>0 then the mean of Y increases as X increases (the mean of Y increases by a multiplicative factor of $\exp(\beta_1)$ per unit of X).
  - If β_1<0 then the mean of Y decreases as X increases (the mean of Y decreases by a multiplicative factor of $\exp(\beta_1)$ per unit of X).

In epidemiology, μ can be thought of as the incidence rate, and $\exp(\beta_1)$ is the **incidence rate ratio.**

31

3. Poisson Regression 32

Recap

- Poisson regression can be used to model data where the outcome is a discrete "count" variable that has a lower limit of 0, but is unlimited in range in the positive direction.
- The model assumes the outcome follows a Poisson distribution
- Since the Poisson distribution approaches a normal distribution as λ becomes large, Poisson regression is the most useful when the mean of the outcome is close to 0
- The Poisson regression approach is also used in epidemiology to study rates of disease occurrence.

32

3. Poisson Regression 33

Recap

- Describe outcomes that would be ideal for Poisson regression
- Compute the probability of observing Y=y given a Poisson distribution with parameter λ

33

Example

Dr. Sangre was examining the factors that related to the number of units of RBC (red blood cells) administered in the operating room during aortic valve surgery. He was interested in whether the new minimally invasive surgery was associated with fewer RBC units, adjusting for patient factors.

Independent Variables
miavr (1=minimally invasive surgery, 0=standard surgery)
agecent (continuous, in years)
white (1=white, 0=otherwise)
male (1=male, 0=female)
hx_db (1=history of diabetes, 0=otherwise)
bmicat (2=30+, 1=25-29.9, 0=<25)

34

How is our outcome variable distributed?

```
> rbcunits %>%
+   select(units) %>%
+   skimr::skim()

-- Variable type: numeric -----------------------------------------------------------------------------------------------
  skim_variable n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100 hist
* <chr>             <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 units                 0             1  1.47  1.90     0     0     1  2.25    10 ▇▂▁▁▁

> rbcunits %>%
+   count(units)
# A tibble: 11 x 2
   units     n
   <dbl> <int>
 1     0   227
 2     1    72
 3     2    70
 4     3    61
 5     4    29
 6     5    15
 7     6     4
 8     7     6
 9     8     3
10     9     2
11    10     3
```

35

Using the calculated mean of 1.47, we can compute the Poisson probability of number of units used.

$$(Y=0)=\frac{e^{-1.47}1.47^{0}}{0!}=0.23$$

$$P(Y=1)=\frac{e^{-1.47}1.47^{1}}{1!}=0.34$$

$$P(Y=10)=\frac{e^{-1.47}1.47^{10}}{10!}=3x10^{-6}$$

36

4. Poisson Regression: An Example

37

How do our independent variables relate to outcome?

```
> rbcunits %>%
+   group_by(miavr) %>%
+   summarise(n = n(), mean = mean(units), sd = sd(units))
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 2 x 4
  miavr     n  mean    sd
  <dbl> <int> <dbl> <dbl>
1     0   198  1.90  2.15
2     1   294  1.18  1.64

> rbcunits %>%
+   group_by(male) %>%
+   summarise(n = n(), mean = mean(units), sd = sd(units))
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 2 x 4
   male     n  mean    sd
  <dbl> <int> <dbl> <dbl>
1     0   199  2.02  1.98
2     1   293  1.10  1.74

> rbcunits %>%
+   group_by(white) %>%
+   summarise(n = n(), mean = mean(units), sd = sd(units))
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 2 x 4
  white     n  mean    sd
  <dbl> <int> <dbl> <dbl>
1     0   125  1.73  2.10
2     1   367  1.38  1.82
```

Given this output, in a Poisson regression analysis, what direction would you expect the beta coefficient for each independent variable to be?

37

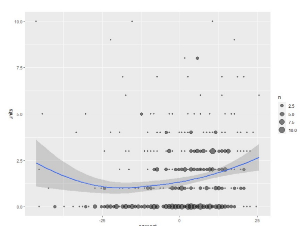
4. Poisson Regression: An Example

38

How do our independent variables relate to outcome?

```
> rbcunits %>%
+   group_by(bmicat) %>%
+   summarise(n = n(), mean = mean(units), sd = sd(units))
# A tibble: 3 x 4
  bmicat     n  mean    sd
   <dbl> <int> <dbl> <dbl>
1      0   157  1.95  1.97
2      1   175  1.04  1.57

> rbcunits %>%
+   group_by(hx_db) %>%
+   summarise(n = n(), mean = mean(units), sd = sd(units))
# A tibble: 2 x 4
  hx_db     n  mean    sd
  <dbl> <int> <dbl> <dbl>
1     0   387  1.40  1.90
2     1   105  1.74  1.87
```



38

4. Poisson Regression: An Example

39

Poisson Regression: No Covariates (Null Model)

```
> summary(rbc0.m)

Call:
glm(formula = units ~ 1, family = "poisson", data = rbcunits)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7155  -1.7155  -0.4129   0.5854   4.6118

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.38631    0.03716   10.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1166.4  on 491  degrees of freedom
Residual deviance: 1166.4  on 491  degrees of freedom
AIC: 1894.6

Number of Fisher Scoring iterations: 5
```

Overall mean # of units = exp(0.3863) = 1.47

Because we didn't include any covariates, this model assumes all patients have the same mean units administered during surgery.

39

Let's add some covariates

Note: Poisson regression allows us to model heterogeneity across patients based on their observed characteristics (independent variables). Each person has their own Poisson mean, based on their X values.

The model we will use is:

$$\ln \mu_{units|X} = \beta_0 + \beta_1 X_{miavr} + \beta_2 X_{agecent} + \beta_3 X_{white} + \beta_4 X_{male} + \beta_5 X_{hxdb} + \beta_6 X_{overwt} + \beta_7 X_{obese}$$

40

Poisson Regression: One Covariate (just to check)

```
> summary(rbc1.m)

Call:
glm(formula = units ~ miavr, family = "poisson", data = rbcunits)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9514  -1.5364  -0.7212   0.6973   5.0097

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.64398    0.05150  12.504  < 2e-16 ***
miavr       -0.47823    0.07439  -6.428 1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1166.4  on 491  degrees of freedom
Residual deviance: 1125.2  on 490  degrees of freedom
AIC: 1855.4

> exp(rbc1.m$coefficients)
(Intercept)       miavr
  1.9040404   0.6198777
```

The rate ratio for a 1-unit increase in "miavr" (having minimally invasive surgery vs. not) is 0.62.

That is, those that have minimally invasive surgery are expected to have 62% the number of RBC units as those who don't.

Those who have minimally invasive surgery are expected to have 38% fewer RBC units compared to those who don't.

41

Poisson Regression: Full Model

```
> summary(rbc2.m)

Call:
glm(formula = units ~ miavr + agecent + white + male + hx_db +
    factor(bmicat), family = "poisson", data = rbcunits)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8197  -1.4244  -0.7190   0.5306   5.1726

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.130836   0.090229  12.533  < 2e-16 ***
miavr           -0.463073   0.076160  -6.080 1.20e-09 ***
agecent          0.010835   0.003077   3.521 0.000429 ***
white           -0.056432   0.084621  -0.667 0.504853
male            -0.500038   0.077011  -6.493 8.41e-11 ***
hx_db            0.179340   0.089241   2.010 0.044473 *
factor(bmicat)1 -0.514948   0.095791  -5.376 7.63e-08 ***
factor(bmicat)2 -0.287145   0.091651  -3.133 0.001730 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 1166.4  on 491  degrees of freedom
Residual deviance: 1006.5  on 484  degrees of freedom
AIC: 1748.7
```

42

4. Poisson Regression: An Example

43

Simplify. Drop white (not significant and does not confound miavr)

```
> summary(rbc3.m)

Call:
glm(formula = units ~ miavr + agecent + male + hx_db + factor(bmicat),
    family = "poisson", data = rbcunits)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8616  -1.4235  -0.6978   0.5361   5.1158

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.10107    0.07887  13.960  < 2e-16 ***
miavr           -0.47226    0.07489  -6.306 2.86e-10 ***
agecent          0.01058    0.00305   3.468 0.000525 ***
male            -0.50586    0.07650  -6.613 3.77e-11 ***
hx_db            0.18522    0.08878   2.086 0.036944 *
factor(bmicat)1 -0.52174    0.09520  -5.481 4.24e-08 ***
factor(bmicat)2 -0.29409    0.09106  -3.230 0.001239 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 1166.4  on 491  degrees of freedom
Residual deviance: 1006.9  on 485  degrees of freedom
AIC: 1747.2
```

43

4. Poisson Regression: An Example

44

Obtain Risk Ratio Values

```
> tibble(
+   parameter = names(rbc3.m$coefficients),
+   rr = exp(rbc3.m$coefficients),
+   as.data.frame.matrix(exp(confint.default(rbc3.m)))
+ )
# A tibble: 7 x 4
  parameter          rr `2.5 %` `97.5 %`
  <chr>           <dbl>   <dbl>    <dbl>
1 (Intercept)     3.01    2.58     3.51
2 miavr           0.624   0.538    0.722
3 agecent         1.01    1.00     1.02
4 male            0.603   0.519    0.701
5 hx_db           1.20    1.01     1.43
6 factor(bmicat)1 0.593   0.492    0.715
7 factor(bmicat)2 0.745   0.623    0.891
```

The mean number of RBC units when all X = 0 (baseline: standard surgery, mean age, female, no diabetes, normal BMI category) is exp(1.101) = 3.01 (95% CI = 2.58, 3.51)

The ratio of mean units RBC for MI vs. standard surgery, controlling for all other variables, is exp(-0.472) = 0.624 (95% CI = 0.538, 0.722).

Can also say: minimally invasive (MI) surgery resulted in a [1-0.62 = 0.38] 38% reduction in the mean number of RBC units (95% CI = 28%, 46%).

44

4. Poisson Regression: An Example

45

Recap

- The Poisson model-building approach is similar to that in other types of regression.
- In Poisson regression, a change in independent variables is associated with a multiplicative change in outcome.

45

4. Poisson Regression: An Example

46

Recap

- Interpret the beta coefficients from a Poisson regression model in terms of the risk ratio

46

5. Poisson Goodness-of-Fit

47

Recall that, in general, **goodness-of-fit** tests always compare the **observed** counts in each category to the **expected** number of counts in each category.

47

5. Poisson Goodness-of-Fit

48

For the **Pearson chi-square goodness of fit** test, we compare the observed counts to the model-predicted counts.

$$\text{Pearson } \chi^2 = \sum_{j=1}^{n} \frac{(y_j - \hat{\mu}_{Y|x})^2}{\hat{\mu}_{Y|x}}$$

This test has df = n – (k+1)
(where k = # independent variables)

For each person, their observed – expected (i.e., the "residual").

48

5. Poisson Goodness-of-Fit

49

Recall the H_0 for the GOF test is "no departure from goodness of fit." Therefore, larger p-values indicate better model fit.

```
> pois_pearson_gof(rbc3.m)
$pval
[1] 1.098586e-52

$df
[1] 485

> pois_dev_gof(rbc3.m)
$pval
[1] 9.445671e-39

$df
[1] 485
```

Yikes—these don't appear to fit that well...

49

5. Poisson Goodness-of-Fit

50

The **deviance chi-square goodness of fit** test compares the model log-likelihood (i.e., deviance) to the maximum possible log-likelihood given the data.

The maximum possible log-likelihood is called the **saturated model**, and contains a separate parameter (μ_i) for each observation i.

This means that, under the saturated model, $\hat{\mu}_i = y_i$.

$$lnL_{maximum} = \sum_{i=1}^{n}(-y_i + y_i \ln(y_i) - \ln(y_i!))$$

Then Deviance $\chi^2 = -2(lnL(model) - lnL(maximum))$ with n-(k+1) df.

Think of it as a model in which each person in the data set has their own dummy variable, and that dummy variable will tell us exactly what their value of Y is.

50

5. Poisson Goodness-of-Fit

51

Recall that comparative fit measures can be used to compare models that may or may not be nested.

They are called "comparative" because their utility comes from the ability to compare different models. However, they are not absolute measures of fit (i.e., you won't get a p-value out of it).

- AIC = Akaike's Information Criterion = -2LL + 2k, k=# parameters
- BIC = Bayesian Information Criterion = -2LL + k(ln(N)), n=# obs.

For both of these, lower values indicate better fit.

There is a penalty for more model parameters, and this penalty is stricter for the BIC.

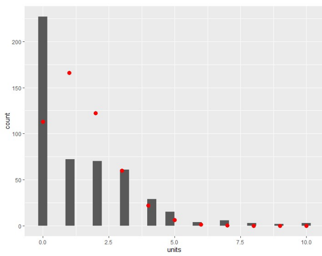
Rule of thumb: For two different models, a difference of AIC or BIC of <3 (or so) indicates no appreciable difference in fit. In most cases, a difference of >10 indicates strong difference in fit.

51

5. Poisson Goodness-of-Fit 52

Poisson Restrictiveness

- In practice, Poisson regression imposes strict assumptions on the distribution of Y
- This is because the Poisson distribution has only one parameter λ, so the mean of Y must equal the variance of Y.
- Other models exist in this situation, such as a **zero-inflated Poisson** model



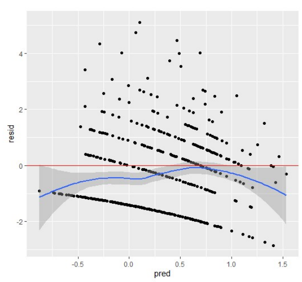
Bars are the actual distribution, while red dots are the expected distribution under Poisson.

52

5. Poisson Goodness-of-Fit 53

Residual Diagnostics

- As anticipated from the previous slide, we appear to be systematically overestimating the counts (the Residuals vs. Fitted plot shows the mean of residuals to be consistently <0 across all predicted values).



53

5. Poisson Goodness-of-Fit 54

Overdispersion

- Recall in Poisson regression $\lambda = \bar{Y} = s_Y^2$
- A quasi-Poisson model allows more flexibility in that $s_y^2 = \tau \cdot \bar{Y}$, where τ is the overdispersion parameter (that tells us whether the variance is larger or smaller than what we would expect under a Poisson distribution).
- We can use the AER package to test $H_0: \tau = 1; H_A: \tau \neq 1$

```
> AER::dispersiontest(rbc3.m)

	Overdispersion test

data:  rbc3.m
z = 5.2797, p-value = 6.469e-08
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  2.300929
```

54

5. Poisson Goodness-of-Fit

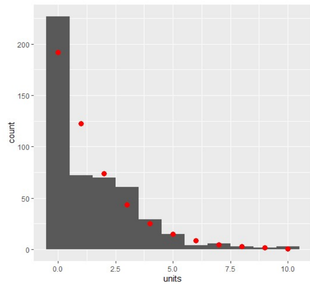
55

Here we see that our outcome of interest in fact is overdispersed with $\tau = 2.30$.

```
> dispersiontest(rbc3.m)

          Overdispersion test

data:  rbc3.m
z = 5.2797, p-value = 6.469e-08
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  2.300929
```



55

5. Poisson Goodness-of-Fit

56

Our original model did not account for the variation (**dispersion**) that exists in the outcome.

What are the consequences of having an overdispersed outcome?

- Smaller estimated standard errors than is realistic
- Smaller p-values
- We will incorrectly find more variables as being statistically significant (higher "false discovery rate")
- Regression coefficients (betas) are appropriately estimated

56

5. Poisson Goodness-of-Fit

57

Recap

- The Poisson GOF or deviance GOF tests can be used to assess the fit of a Poisson regression model.
- In practice, the Poisson model is rarely fit well as it assumes the mean is equal to the variance.
- One way to accommodate this strict assumption is to fit a quasi-Poisson model, which allows for a different value of the variance.

57

5. Poisson Goodness-of-Fit 58

Recap

- Diagnose the fit of a Poisson regression model
- Detect and explain the consequences of Poisson overdispersion

58

6. Negative Binomial Regression 59

We can use the **negative binomial distribution** to account for overdispersed outcome variables.

A little bit about this distribution:

- Used to model the number of failures in a series of Bernoulli trials until a success is observed (e.g., how many times would you have to flip a coin until it came up heads?)
- Conditional on the mean, the random variable Y is distributed as Poisson.
- The mean is a function of the gamma distribution with shape parameter *k*
- Gamma distributions are a family of probability distributions for continuous random variables, defined by both a scale and shape parameter.

59

6. Negative Binomial Regression 60

For count data (y=0, 1, 2, etc.) the negative binomial distribution function is:

$$P(Y = y \mid \mu, k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{k}{\mu+k}\right)^{k}\left(1-\frac{k}{\mu+k}\right)^{y}$$

To simplify with interpretation later, we'll let $\alpha = 1/k$.

60

6. Negative Binomial Regression

For the variable in question:

E(Y) = μ

Var(Y) = μ + $\boldsymbol{\alpha\mu^2}$

Note a couple things about this:

- Without the "+ $\alpha\mu^2$" part, the expected value and variance would be the same as a Poisson variable.
- The α term explicitly models the overdispersion (i.e., "extra-Poisson variation")
- The α term is assumed constant over all values of X.
- In R, there is a different notation: Var(Y) = $\mu + \frac{\mu^2}{\theta}$. Therefore, $\theta = \frac{1}{\alpha}$.

What is the value of θ when there is no overdispersion?

61

6. Negative Binomial Regression

Let's estimate the RBC model with negative binomial regression instead:

```
> summary(rbc4.m)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9267  -1.1883  -0.4698   0.3278   2.7875

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.200270   0.133720   8.976  < 2e-16 ***
miavr           -0.536326   0.117634  -4.559 5.13e-06 ***
agecent          0.010003   0.004573   2.187  0.02873 *
male            -0.561477   0.119174  -4.711 2.46e-06 ***
hx_db            0.246874   0.143282   1.723  0.08489 .
factor(bmicat)1 -0.569508   0.144816  -3.933 8.40e-05 ***
factor(bmicat)2 -0.400984   0.147295  -2.722  0.00648 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.1149) family taken to be 1)

    Null deviance: 589.06  on 491  degrees of freedom
Residual deviance: 515.29  on 485  degrees of freedom
AIC: 1587

Number of Fisher Scoring iterations: 1

              Theta:  1.115
          Std. Err.:  0.161

 2 x log-likelihood:  -1571.040
```

1. Note the p-values have changed from the Poisson model.

2. Note the deviance is much lower compared to the Poisson model.

3. Our estimate of θ.

62

6. Negative Binomial Regression

The model now fits better:

```
> pois_pearson_gof(rbc4.m)
$pval
[1] 4.023208e-54

$df
[1] 485

> pois_dev_gof(rbc4.m)
$pval
[1] 0.1649212

$df
[1] 485
```

The Pearson chi-square test will reject H_0 quite frequently when the sample size is this large.

The Deviance GOF doesn't show departure from good fit.

63

6. Negative Binomial Regression 64

Recap

- For almost all purposes, negative binomial regression is treated the same as Poisson regression
- Negative binomial regression explicitly models the overdispersion in the dependent variable.

64

6. Negative Binomial Regression 65

Recap

- ➢ Explain when it is appropriate to use negative binomial regression
- ➢ Fit a negative binomial regression model
- ➢ Diagnose the overdispersion present in the outcome by interpreting the output from a negative binomial model

65

7. Rate Outcomes 66

What is a **rate**?

A rate is just a **count** divided by some population denominator.

E.g.,

- Number of unemployment claims *per 100 people in the state.*
- Number of automobile deaths *per 10,000 truck miles traveled.*
- Number of false start penalties *per minute of football played.*
- Number of people signing a petition *per 1000 people solicited.*

A rate is a way of making a count **comparable** across different-sized populations.

66

Example 2

A case management program for depression was tested in a local hospital that cares for the indigent and homeless, who often access health care by arriving in the emergency room.

Investigators wanted to know whether implementation of the new program reduced the number of times individuals visited the ER.

67

Y = # of ER visits in the year following treatment for depression

TRT = treatment group (0=usual care, 1=new program)

Investigators noted that ER visits vary greatly depending on whether the individual uses alcohol or IV drugs. They wanted to control for:

RACE (0=white, 1=non-white)

ALC (continuous measure of alcohol use)

DRUG (continuous measure of IV drug use)

68

In this dataset we observed the following:

- 1/3 of subjects had Y=0 (no ER visits within one year)
- ½ had either Y=0 or Y=1.

This means the **event is rare**. Since it is unlikely to occur, the number of observed counts is low.

For rare events, the Poisson distribution is strongly skewed with many 0 and 1 values.

69

7. Rate Outcomes

70

In OLS regression, if the outcome is heavily skewed then we can apply a **transformation** to it.

In this case, the variable is highly non-normal and cannot be transformed to normality using a natural log (or other) transformation.

Fortunately, a Poisson model is a good way to model this type of data.

70

7. Rate Outcomes

71

In this example, our Poisson regression equation for the number of ER visits in the year following treatment is given by:

$$\ln\big(E(Y_i)\big) = \beta_0 + \beta_1 TRT_i + \beta_2 RACE_i + \beta_3 DRG_i + \beta_4 ALC_i$$

Where $Y_i \sim$ Poisson(μ_i)

This is a model that we know how to implement.

71

7. Rate Outcomes

72

In this example, we assume that we have followed individuals for one year to track their ER visits.

However, what if in an individual was followed for only half a year?

E.g., Person A had 2 ER visits in the past year. Person B had 1 ER visit, but was only followed for half a year.

Person A's rate is $\frac{2\ ER\ visits}{1\ year} = 2$ visits/year

Person B's rate is $\frac{1\ ER\ visit}{1/2\ year} = 2$ visits/year

72

7. Rate Outcomes

73

One way to accommodate differences among subjects with regard to follow-up time is to model the mean count per unit time.

We typically model the expected number of counts $E(Y_i)$

However, we can also model the expected rate $E(Y_i)/\boldsymbol{t_i}$

How would this change our regression equation?

73

7. Rate Outcomes

74

Rate: $\ln(E(Y_i)/\boldsymbol{t_i}) = \beta_0 + \beta_1 TRT_i + \beta_2 RACE_i + \beta_3 DRG_i + \beta_4 ALC_i$

$$\ln\big(E(Y_i)\big) - \mathbf{ln}(\boldsymbol{t_i}) = \beta_0 + \beta_1 TRT_i + \beta_2 RACE_i + \beta_3 DRG_i + \beta_4 ALC_i$$

Count: $\ln\big(E(Y_i)\big) = \beta_0 + \beta_1 TRT_i + \beta_2 RACE_i + \beta_3 DRG_i + \beta_4 ALC_i + \mathbf{ln}(\boldsymbol{t_i})$

If we model a rate…

…it's really just a Poisson count model…

…with this extra term at the end.

74

7. Rate Outcomes

75

$$\ln\big(E(Y_i)\big) = \beta_0 + \beta_1 TRT_i + \beta_2 RACE_i + \beta_3 DRG_i + \beta_4 ALC_i + \mathbf{ln}(\boldsymbol{t_i})$$

What is this extra term?

- It **doesn't have a beta coefficient** to estimate; the term associated with it is fixed to 1.
- This term accounts for follow-up time, and is called an **offset**.

75

We can also use this offset to account for having different maximum possible counts.

- If the count is number of games won, the offset could be the number of games played.
- If the count is number of individuals who voted, the offset could be the total population of individuals under consideration.

76

Why do we have to use the offset? Can't we just calculate the rate and then use it directly as the dependent variable?

No!

- Remember, everything we've discussed about Poisson regression so far requires that the outcome be a discrete count variable, not a continuous rate.
- Furthermore, modelling the offset is mathematically equivalent to modeling a rate.

77

Example 3 (Agresti)

Suppose we want to model whether accidents at road/train crossings has been increasing over time.

Our observations are number of accidents at the year level.

However, there is more train activity in some years, so we want to include an offset for the total km (in millions) in train travel in that year.

❑ Should we include km as a covariate or an offset?

Since km of train travel sets a limit to the number of accidents that could happen, we should model it as an offset.

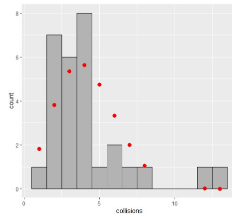
78

7. Rate Outcomes

Our equation for this model is:

$$\ln(E(\#\ collisions_i)/\boldsymbol{km_i}) = \beta_0 + \beta_1(time)_i$$

$$\ln\big(E(\#\ collisions_i)\big) = \beta_0 + \beta_1(time)_i + \mathbf{ln}(\boldsymbol{km_i})$$



79

7. Rate Outcomes

Let's include log(km) as the offset:

```
> summary(rr1.m)

Call:
glm(formula = collisions ~ time.1975 + offset(log(km)),
 family = poisson, data = railroad)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0580  -0.7825  -0.0826   0.3775   3.3873

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.21142    0.15892  -26.50  < 2e-16 ***
time.1975   -0.03292    0.01076   -3.06  0.00222 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 47.376  on 28  degrees of freedom
Residual deviance: 37.853  on 27  degrees of freedom
AIC: 133.52

> glm.RR(rr1.m, digits=4)
Waiting for profiling to be done...
                RR  2.5 % 97.5 %
(Intercept) 0.0148 0.0107 0.0200
time.1975   0.9676 0.9472 0.9881
```

Our fit equation is:

Estimated number of accidents.

$$\ln\big(E(\#\ collisions_i)\big) = -4.21142 - 0.03292(time)_i + \mathbf{ln}(\boldsymbol{km_i})$$

$$E(\#\ collisions_i) = (\boldsymbol{km_i})e^{-4.211-0.033(time)_i}$$

Estimated number of accidents **per million km**.

$$\ln(E(\#\ collisions_i)/\boldsymbol{km_i}) = -4.21142 - 0.03292(time)_i$$

$$E(\#\ collisions_i)/\boldsymbol{km_i} = e^{-4.211\quad .033(time)_i}$$

Estimated number of accidents per million km when time.1975 = 0 (i.e., in 1975).

Each additional year is associated with 0.968 times the rate of accidents (p=.002).

80

7. Rate Outcomes

The number of accidents for a given value of km is:

$$\ln\big(E(\#\ collisions_i)\big) = -4.211418 - 0.032918(time)_i + \mathbf{ln}(\boldsymbol{km_i})$$

$$E(\#\ collisions_i) = (\boldsymbol{km_i})e^{-4.211-0.033(time)_i}$$

The number of accidents per million km is:

$$\ln(E(\#\ collisions_i)/\boldsymbol{km_i}) = -4.211418 - 0.032918(time)_i$$

$$E(\#\ collisions_i)/\boldsymbol{km_i} = e^{-4.211-0.033(time)_i}$$

81

7. Rate Outcomes

The predict() function will automatically predict ln(E(Y)).

- It uses information about X and the offset to make predictions.
- To get the predicted *rate*, you must feed predict() an offset variable of 1 values.

```
railroad %>%
  mutate(pred_rate = predict(rr1.m,
                             tibble(
                               time.1975 = railroad$time.1975,
                               km = 1),
                             type = "response"),
         pred_count = predict(rr1.m, ., type = "response")
  )
# A tibble: 29 x 6
   time    km collisions time.1975 pred_rate pred_count
  <dbl> <dbl>      <dbl>     <dbl>     <dbl>      <dbl>
1  2003   518          3        28   0.00590       3.06
2  2002   516          3        27   0.00610       3.15
3  2001   508          4        26   0.00630       3.20
4  2000   503          3        25   0.00651       3.27
```

This is the predicted units of Y *per unit offset* (i.e., per million km)

This is the predicted units of Y, taking the offset into account.

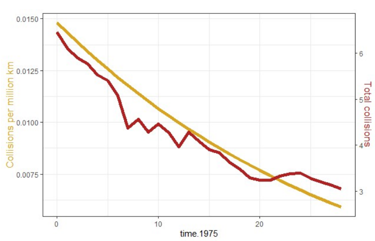
$\exp(-4.21 - 0.033(28) + \ln(518))$

$\exp(-4.21 - 0.033(28))$

82

7. Rate Outcomes

Note that the predicted rates are modeled as an exponential trend over time, but there are some "hiccups" with the counts. These are probably in years when the total km of train traffic was particularly high or low.



83

7. Rate Outcomes

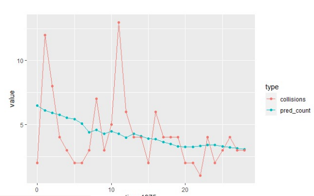
Check the GOF of the model.

```
> pois_pearson_gof(rr1.m)
$pval
[1] 0.03151

$df
[1] 27

> pois_dev_gof(rr1.m)
$pval
[1] 0.08021703

$df
[1] 27
```



The observed counts in any given year are sometimes vastly different from the predicted counts.

What could this be saying about our model?

84

7. Rate Outcomes 85

Recap

- Instead of modeling a count outcome with Poisson regression, we can model a rate.
- A rate is any variable that has in its denominator information about the maximum possible counts.
- Rates cannot be modeled directly; they need to be included as an offset in Poisson regression.

85

7. Rate Outcomes 86

Recap

- ➢ Determine when it is appropriate to model the outcome as a rate
- ➢ Implement the analysis of a rate outcome by the inclusion of an offset term
- ➢ When using an offset, interpret the regression output with regard to either the count or rate of outcome

86

8. Recap 87

Summary

- Poisson regression is a way to model discrete count data, or rate data (counts per some denominator unit)
- The Poisson assumption isn't easily satisfied and can often be overdispersed
- Overdispersion may be caused by excess zeroes, in which case a zero-inflated model may be better
- Negative Binomial models allow for an overdispersion parameter in the modeling approach
- Modeling a rate through Poisson or NB regression will require an "offset" term

87

8. Recap 88

Additional Reading

- Commentary on the Deviance GOF: https://thestatsgeek.com/2014/04/26/deviance-goodness-of-fit-test-for-poisson-regression/
- Examining Poisson overdispersion: http://biometry.github.io/APES//LectureNotes/2016-JAGS/Overdispersion/OverdispersionJAGS.html
- Zero-Inflated Poisson Regression: https://stats.idre.ucla.edu/r/dae/zip/

88

6. Recap 89

Packages and Functions

- `AER::dispersiontest()`
- `MASS::glm.nb()`

89