BIOS 6312: Modern Regression Analysis

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 8: Regression with Discrete Outcomes (Non-binary outcomes)

Version: 03/09/2023

Table of Contents

1. Discrete outcomes
2. Regression of nominal outcomes
3. Regression of ordinal outcomes
4. Poisson regression (count outcomes)

Discrete outcomes

Basic ideas:

- In the previous set of notes, all outcomes were binary (e.g., CHD, esophageal cancer).
- In this set of notes, we'll tackle three additional kinds of discrete outcomes:
  - Nominal outcomes (categorical, unordered).
  - Ordinal outcomes (categorical, ordered).
  - Count outcomes (non-negative integers).
- We'll pay particular attention to the challenges in interpretation that arise in each of these three settings.

Table of Contents

1. Discrete outcomes
2. Regression of nominal outcomes
3. Regression of ordinal outcomes
4. Poisson regression (count outcomes)

Nominal outcomes

Example: Diabetes and CHD in MRI cohort

- Recall the motivating example from the MRI cohort:
  - X: $0 =$ no diabetes; $1 =$ diabetes.
  - Y: $0 =$ no CHD; $1 =$ angina/myocardial infarction.

	CHD	No CHD	Total
Diabetes	23	56	79
No diabetes	132	524	656
Total	155	580	735

- This example relied on our ability to dichotomize CHD (something that is not always possible—and frankly, not always desirable).

Nominal outcomes

Example: Diabetes and CHD in MRI cohort

- Here is more complete representation of the same data.
  - X: 0 = no diabetes; 1 = no diabetes.
  - Y: 0 = no CHD; 1 = angina; 2 = myocardial infarction.

	MI	Angina	No CHD	Total
Diabetes	16	7	56	79
No diabetes	75	57	524	656
Total	91	64	580	735

- The `tabulate` command in Stata will arrange the variables in a different order.

Nominal outcomes

Setup: Nominal outcomes

- A nominal outcome with M unordered categories $(1, \ldots, M)$ follows a categorical distribution (special case of the multinomial distribution).
  - $Y \sim \text{Multinomial}(1, \mathbf{p})$; $\mathbf{p} = (p_1, \ldots, p_M)$, $\sum_{m=1}^{M} p_m = 1, p_m > 0$.
  - $\text{P}(Y = m) = p_m$.
  - You can think of a realization of Y as a single number from 1 to M, or you can think of it as a vector of all zeros except a one in the position corresponding to its realization. Because the categories are unordered, it is easiest to think of it in the latter way.
    - ⋆ $\text{E}[Y] = \mathbf{p}$.
    - ⋆ $\text{Var}[Y] = \text{diag}_M(\mathbf{p}(\mathbf{1} - \mathbf{p})^T + \mathbf{p}\mathbf{p}^T) - \mathbf{p}\mathbf{p}^T$.
  - You do not need to remember these formulas in this class.
- Next goal: Regression with nominal outcomes.

Multinomial regression

Regression of categorical outcomes:

- Y has M levels, $1, \ldots, M$.
- Choose $Y = M$ as reference category without loss of generality.
- Consider a predictor X:

$$\begin{aligned}
\log\left(\frac{\mathrm{P}(Y=1|X=x)}{\mathrm{P}(Y=M|X=x)}\right) &= \beta_{01} + \beta_{11}x \\
\log\left(\frac{\mathrm{P}(Y=2|X=x)}{\mathrm{P}(Y=M|X=x)}\right) &= \beta_{02} + \beta_{12}x \\
&\vdots \\
\log\left(\frac{\mathrm{P}(Y=M-1|X=x)}{\mathrm{P}(Y=M|X=x)}\right) &= \beta_{0(M-1)} + \beta_{1(M-1)}x
\end{aligned}$$

- Resembles $M-1$ logistic models (common ref. category).

Multinomial regression

Regression of categorical outcomes:

- Model: $\log\left(\frac{P(Y=m|X=x)}{P(Y=M|X=x)}\right) = \beta_{0m} + \beta_{1m}x,$ for $m = 1, \ldots, M-1$.
- Re-expressing:

$$\begin{aligned}
P(Y=1|X=x) &= P(Y=M|X=x)\exp(\beta_{01} + \beta_{11}x) \\
P(Y=2|X=x) &= P(Y=M|X=x)\exp(\beta_{02} + \beta_{12}x) \\
&\vdots \\
P(Y=M-1|X=x) &= P(Y=M|X=x)\exp(\beta_{0(M-1)} + \beta_{1(M-1)})
\end{aligned}$$

- The M^{th} category is implied:

$$\begin{aligned}
P(Y=M|X=x) &= 1 - \sum_{m=1}^{M-1} P(Y=m|X=x) \\
&\Rightarrow \quad P(Y=M|X=x) = \frac{1}{1 + \sum_{m=1}^{M-1}\exp(\beta_{0m} + \beta_{1m}x)}
\end{aligned}$$

Multinomial regression

Regression of categorical outcomes:

- Re-expressing (again), note that for $m = 1, \ldots, M-1$:

$$P(Y = m|X = x) \quad = \quad \frac{\exp(\beta_{0m} + \beta_{1m}x)}{1 + \sum_{j=1}^{M-1} \exp(\beta_{0j} + \beta_{1j}x)}$$

- $\exp(\beta_{1m})$: "ratio of risk ratios"—comparing relative proportion of $Y = m$ to $Y = M$ between subgroups differing in X by one unit.
- Typically, estimation is performed by *maximum a posteriori* (MAP) estimation, a method we won't get into in this course but one you can look up if you're interested :).

Multinomial regression

Example: Diabetes and CHD in MRI cohort

- Let us use the MRI study to re-examine the association between diabetes and coronary heart disease.
  - X: $0 =$ no diabetes; $1 =$ diabetes.
  - Y: $0 =$ no CHD; $1 =$ angina; $2 =$ myocardial infarction.
    - Note: We do *not* dichotomize Y in this example!
- Model: $\log\left(\frac{\mathrm{P}(Y=m|X=x)}{\mathrm{P}(Y=0|X=x)}\right) = \beta_{0m} + \beta_{1m}x$, for $m = 1, 2$.
  - Example: $\exp(\beta_{01})$ denotes the prevalence ratio $\frac{\mathrm{P}(Y=1|X=0)}{\mathrm{P}(Y=0|X=0)}$.
  - Example: $\exp(\beta_{12})$ denotes a ratio of prevalence ratios:

$$\frac{\mathrm{P}(Y=2|X=1)/\mathrm{P}(Y=0|X=1)}{\mathrm{P}(Y=2|X=0)/\mathrm{P}(Y=0|X=0)}.$$

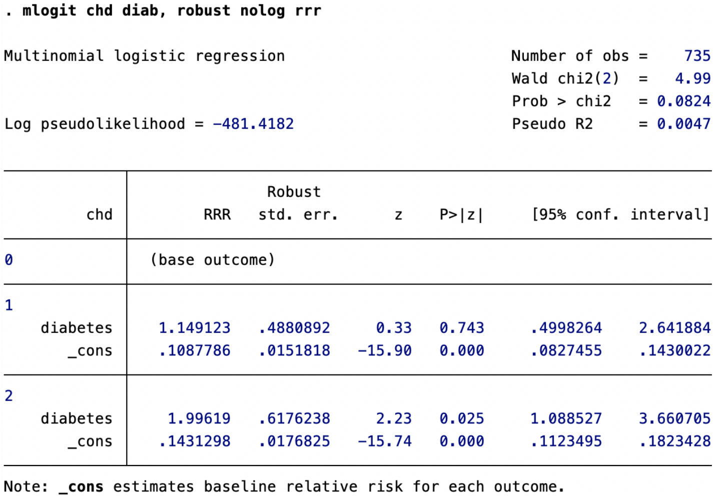
Multinomial regression

Example: Diabetes and CHD in MRI cohort

- Variables:
  - X: $0 =$ no diabetes; $1 =$ diabetes.
  - Y: $0 =$ no CHD; $1 =$ angina; $2 =$ myocardial infarction.
- Model: $\log\left(\frac{\mathrm{P}(Y=m|X=x)}{\mathrm{P}(Y=0|X=x)}\right) = \beta_{0m} + \beta_{1m}x,$ for $m = 1, 2$.
- Multinomial regression in Stata: `mlogit`.
  - Option `rrr` exponentiates to provide the relative risk ratios.
  - Option `baseoutcome()` allows you to set reference group.
  - Option `robust` provides sandwich variance.

Multinomial regression

Example: Diabetes and CHD in MRI cohort



```
. mlogit chd diab, robust nolog rrr

Multinomial logistic regression                         Number of obs =    735
                                                        Wald chi2(2)  =   4.99
                                                        Prob > chi2   = 0.0824
Log pseudolikelihood = -481.4182                        Pseudo R2     = 0.0047

------------------------------------------------------------------------------
             |               Robust
         chd |        RRR   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
0            |  (base outcome)
-------------+----------------------------------------------------------------
1            |
    diabetes |   1.149123   .4880892     0.33   0.743     .4998264    2.641884
       _cons |   .1087786   .0151818   -15.90   0.000     .0827455    .1430022
-------------+----------------------------------------------------------------
2            |
    diabetes |    1.99619   .6176238     2.23   0.025     1.088527    3.660705
       _cons |   .1431298   .0176825   -15.74   0.000     .1123495    .1823428
------------------------------------------------------------------------------
Note: _cons estimates baseline relative risk for each outcome.
```

Multinomial regression

Example: Diabetes and CHD in MRI cohort

- Resembles output from two logistic regression models!
- As examples, let's work through the following exercises:
  1. Determine whether there is evidence of an overall association between diabetes and CHD category.
  2. Compare the prevalence ratio (comparing the prevalence of MI to that of no CHD) between those with and without diabetes.
  3. Compare the prevalence ratio (comparing the prevalence of MI to that of angina) between those with and without diabetes.

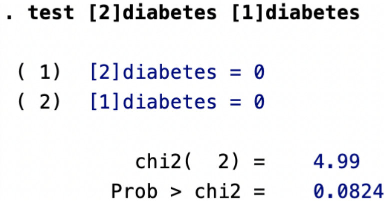
Multinomial regression

Example: Diabetes and CHD in MRI cohort

- **Example 1**: Determine whether there is evidence of an overall association between diabetes and CHD category.
  - No overall association between diabetes and CHD category implies that both $\beta_{11} = 0$ and $\beta_{12} = 0$.
  - We can use the `test` command; however, we need to use some Stata-specific notation to account for the fact that the parameters are coming from different parts of the model: `test [2]diabetes [1]diabetes`.
  - The `testparm` command can also be used in this case and doesn't require the same specific notation.

Multinomial regression

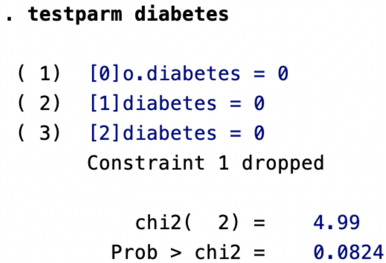
Example: Diabetes and CHD in MRI cohort



```
. test [2]diabetes [1]diabetes

 ( 1)  [2]diabetes = 0
 ( 2)  [1]diabetes = 0

           chi2(  2) =    4.99
         Prob > chi2 =    0.0824
```

Multinomial regression

Example: Diabetes and CHD in MRI cohort



```
. testparm diabetes

 ( 1)  [0]o.diabetes = 0
 ( 2)  [1]diabetes = 0
 ( 3)  [2]diabetes = 0
       Constraint 1 dropped

           chi2(  2) =    4.99
         Prob > chi2 =    0.0824
```

Multinomial regression

Example: Diabetes and CHD in MRI cohort

- **Example 2**: Compare the prevalence ratio (comparing the prevalence of MI to that of no CHD) between those with and without diabetes.
  - This information is available to us in the output.
  - From this model, we see evidence that the prevalence ratio (comparing the prevalence of MI to that of no CHD) differs between those with and without diabetes (RRR=1.996; 95% CI: [1.0885, 3.66]; p=0.025).

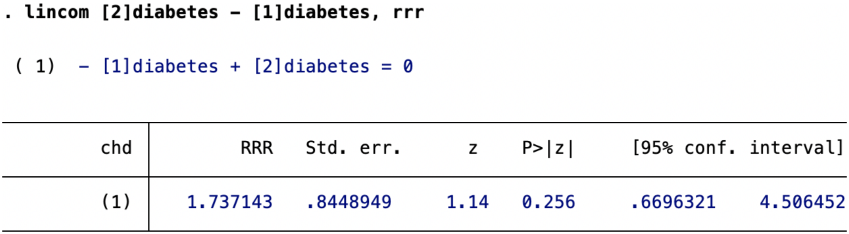
Multinomial regression

Example: Diabetes and CHD in MRI cohort

- **Example 3**: Compare the prevalence ratio (comparing the prevalence of MI to that of angina) between those with and without diabetes.
  - This information is not available to us in the output, but is encoded in the model.
  - With a little regression math (that I will leave to you as an exercise), we see that the RRR is represented by $\exp(\beta_{12} - \beta_{11})$.
  - It should therefore come as little surprise that the `lincom` command will be useful. The same notation used for the `test` command carries over, and the `rrr` option is necessary if it was not included in the `mlogit` command.

Multinomial regression

Example: Diabetes and CHD in MRI cohort



```
. lincom [2]diabetes - [1]diabetes, rrr

 ( 1)  - [1]diabetes + [2]diabetes = 0

------------------------------------------------------------------------------
         chd |        RRR   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         (1) |   1.737143   .8448949     1.14   0.256     .6696321    4.506452
------------------------------------------------------------------------------
```

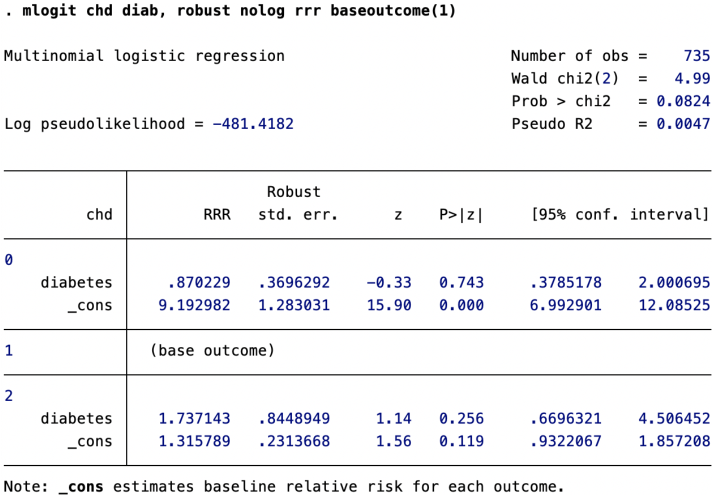
Multinomial regression

Example: Diabetes and CHD in MRI cohort

- We could also tell Stata to set the base outcome in order to learn the same information. Since we are comparing MI ($Y = 2$) to angina ($Y = 1$), we should see these numbers show up by setting the base outcome to 1.

Multinomial regression

Example: Diabetes and CHD in MRI cohort



```
. mlogit chd diab, robust nolog rrr baseoutcome(1)

Multinomial logistic regression                         Number of obs =    735
                                                        Wald chi2(2)  =   4.99
                                                        Prob > chi2   = 0.0824
Log pseudolikelihood = -481.4182                        Pseudo R2     = 0.0047

------------------------------------------------------------------------------
             |               Robust
         chd |        RRR   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
0            |
    diabetes |    .870229   .3696292    -0.33   0.743     .3785178    2.000695
       _cons |   9.192982   1.283031    15.90   0.000     6.992901    12.08525
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
    diabetes |   1.737143   .8448949     1.14   0.256     .6696321    4.506452
       _cons |   1.315789   .2313668     1.56   0.119     .9322067    1.857208
------------------------------------------------------------------------------
Note: _cons estimates baseline relative risk for each outcome.
```

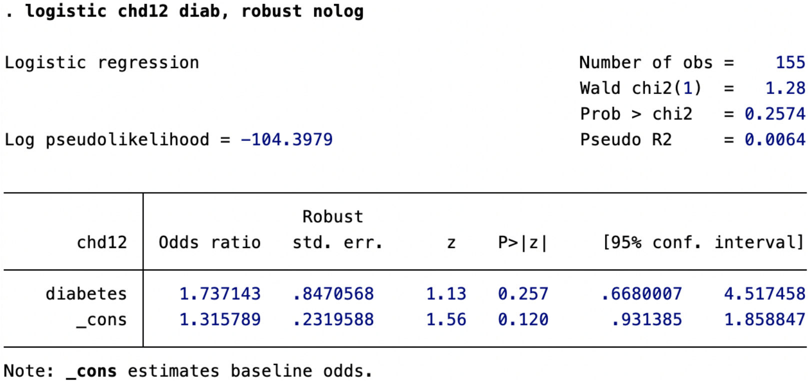
Multinomial regression

Example: Diabetes and CHD in MRI cohort

- This is a saturated model. Therefore, we would *further* expect the model to agree with a model that split this problem up into separate logistic models.
  - It does, with a small caveat, which we will now illustrate.
- Let us repeat Example 3 by re-coding the variables as follows:
  - X: $0 =$ no diabetes; $1 =$ diabetes.
  - Y: $0 =$ angina, $1 =$ myocardial infarction.
    - `gen chd12 = chd - 1`
    - `replace chd12 = . if chd == 0`
- Model: $\text{logit}(\text{P}(Y = 1|X = x)) = \beta_0 + \beta_1 x$.
  - $\exp(\beta_1)$ represents the ratio of prevalence ratios described in Example 3 on the previous slide.

Multinomial regression

Example: Diabetes and CHD in MRI cohort



```
. logistic chd12 diab, robust nolog

Logistic regression                                     Number of obs =    155
                                                        Wald chi2(1)  =   1.28
                                                        Prob > chi2   = 0.2574
Log pseudolikelihood = -104.3979                        Pseudo R2     = 0.0064

------------------------------------------------------------------------------
             |               Robust
       chd12 | Odds ratio   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    diabetes |   1.737143   .8470568     1.13   0.257     .6680007    4.517458
       _cons |   1.315789   .2319588     1.56   0.120      .931385    1.858847
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Multinomial regression

Example: Diabetes and CHD in MRI cohort

- Multinomial logit model on whole sample:
  - RRR=1.737
  - 95% CI: [0.670, 4.51]
  - p=0.256
- Logistic model on subset:
  - RRR=1.737
  - 95% CI: [0.668, 4.52]
  - p=0.257
- The point estimates are identical, as we might expect.
- The standard errors are being computed a little bit differently (this is reflected in the degrees of freedom, for instance). Asymptotically, they will agree (and the finite-sample discrepancies are typically negligible).

Multinomial regression

Additional thoughts:

- Multinomial regression essentially handles categorical outcomes by splitting it into several logistic regression problems (but of course, the separate models are estimated simultaneously).
- Needless to say, you should be able to generalize the ideas of adjustment, interactions, categorical covariates, splines, transformations, etc. to models involving nominal outcomes.

Table of Contents

1. Discrete outcomes
2. Regression of nominal outcomes
3. Regression of ordinal outcomes
4. Poisson regression (count outcomes)

Regression of ordinal outcomes

Ordinal variables:

- For cases in which Y has a clear ordering, we may be comfortable with a simplifying assumption regarding the odds ratios.
- Again let X denote our predictor of interest, and suppose that Y has M *ordered* categories, $1, \ldots, M$.
- Ordered logit model (proportional odds):

$$\operatorname{logit}\left(\mathrm{P}(Y \leq m|X = x)\right) = \log\left(\frac{\mathrm{P}(Y \leq m|X = x)}{\mathrm{P}(Y > m|X = x)}\right) = \beta_{0m} - \beta x$$

- Each of the $M - 1$ models gets its own intercept, but the coefficient corresponding to X is shared.
- Because of how the model is parameterized, we need to be careful in our interpretation.

Regression of ordinal outcomes

Proportional odds regression:

- Ordered logit model:

$$\log\left(\frac{\mathrm{P}(Y \leq m|X = x)}{\mathrm{P}(Y > m|X = x)}\right) = \beta_{0m} - \beta x$$

- Let's start with the baseline odds:
  - $\exp(\beta_{01}) = O(Y = 1|X = 0)$.
  - $\exp(\beta_{02}) = O(Y \in \{1, 2\}|X = 0)$.
  - $\vdots$
  - $\exp(\beta_{0(M-1)}) = O(Y \in \{1, 2, \ldots, M - 1\}|X = 0)$.

Regression of ordinal outcomes

Proportional odds regression:

- To interpret β, note that for $m = 1, \ldots, M-1$:

$$
\begin{aligned}
& \mathrm{logit}\left(\mathrm{P}(Y \le m | X = x+1)\right) - \mathrm{logit}\left(\mathrm{P}(Y \le m | X = x)\right) \\
= \quad & (\beta_{0m} - \beta(x+1)) - (\beta_{0m} - \beta x) = -\beta \\
\Rightarrow \quad & \exp(-\beta) = \frac{\mathrm{P}(Y \le m | X = x+1)/\mathrm{P}(Y > m | X = x+1)}{\mathrm{P}(Y \le m | X = x)/\mathrm{P}(Y > m | X = x)} \\
= \quad & \frac{O(Y \le m | X = x+1)}{O(Y \le m | X = x)} \\
\Rightarrow \quad & \exp(\beta) = \frac{O(Y > m | X = x+1)}{O(Y > m | X = x)}
\end{aligned}
$$

Regression of ordinal outcomes

Proportional odds regression:

- After all the math, we have the following expression:

$$\exp(\beta) = \frac{O(Y > m|X = x + 1)}{O(Y > m|X = x)}$$

- Truthfully, this is a difficult parameter to wrap your mind around.
- Common interpretation: the subgroup $X = x + 1$ has $\exp(\beta)$ times the odds of "being in a higher category" as compared to those with $X = x$. This interpretation doesn't seem to follow, as the category being compared is the same for both subgroups.
- Better interpretation: for each $m = 1, \ldots, M - 1$, the odds of Y exceeding m for the subgroup $X = x + 1$ is $\exp(\beta)$ times that of the subgroup $X = x$. This tricky interpretation is a consequence of using the ordered logit rather than, say, an "adjacent categories" logit.

Regression of ordinal outcomes

Example of proportional odds:

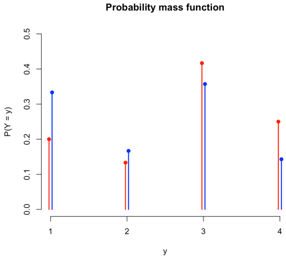
- Take the following table as an example:

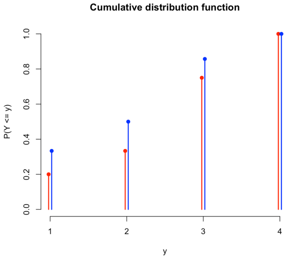
	$x = 0$	$x = 1$
$\mathrm{P}(Y = 1 \mid X = x)$	1/5	1/3
$\mathrm{P}(Y = 2 \mid X = x)$	2/15	1/6
$\mathrm{P}(Y = 3 \mid X = x)$	5/12	5/14
$\mathrm{P}(Y = 4 \mid X = x)$	1/4	1/7

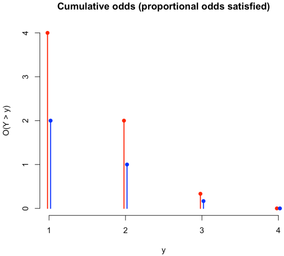
- That the proportional odds assumption is met is not even close to obvious. Nevertheless, if we determine $O(Y > j \mid X = x)$ for each j and each x, it becomes a little clearer:

	$x = 0$	$x = 1$	OR
$O(Y > 1 \mid X = x)$	4	2	1/2
$O(Y > 2 \mid X = x)$	2	1	1/2
$O(Y > 3 \mid X = x)$	1/3	1/6	1/2

Example of proportional odds:


Probability mass function


Cumulative distribution function


Cumulative odds (proportional odds satisfied)

Regression of ordinal outcomes

Example of proportional odds:

- Table of odds for each group:

	$x = 0$	$x = 1$	OR
$O(Y > 1 \| X = x)$	4	2	$1/2$
$O(Y > 2 \| X = x)$	2	1	$1/2$
$O(Y > 3 \| X = x)$	$1/3$	$1/6$	$1/2$

- From this table, we can actually write down the parameters of the proportional odds model. The "intercepts" are setting the distribution for $X = 0$, and then the single parameter β is telling you how to jump from $X = 0$ to $X = 1$:
  - $\beta_{01} = \text{logit}(\text{P}(Y \leq 1|X = 0)) = \log(1/4)$.
  - $\beta_{02} = \text{logit}(\text{P}(Y \leq 2|X = 0)) = \log(1/2)$.
  - $\beta_{03} = \text{logit}(\text{P}(Y \leq 3|X = 0)) = \log(3)$.
  - $\beta = \log(1/2)$.

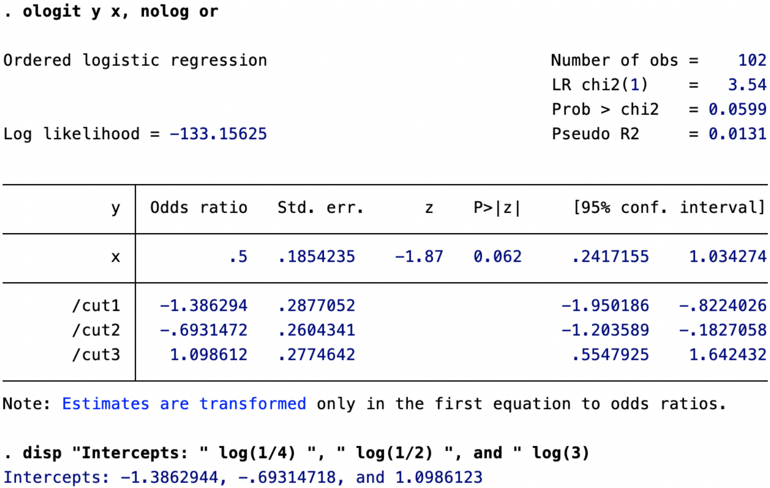
Regression of ordinal outcomes

Example of proportional odds:

- To construct a data set with a joint distribution of (X, Y) as per our example table, we need 102 observations ($N_0 = 60$ and $N_1 = 42$).
- In Stata, the proportional odds cumulative logit model can be fit using the `ologit` command.
  - Option `robust` provides sandwich variance (although that is not the point of this particular example).
  - Option `nolog` suppresses iterative output.
  - Option `or` exponentiates to provide the odds ratios.

Regression of ordinal outcomes

Example of proportional odds:



```
. ologit y x, nolog or

Ordered logistic regression                             Number of obs =    102
                                                        LR chi2(1)    =   3.54
                                                        Prob > chi2   = 0.0599
Log likelihood = -133.15625                             Pseudo R2     = 0.0131

------------------------------------------------------------------------------
           y | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
           x |         .5   .1854235    -1.87   0.062     .2417155    1.034274
-------------+----------------------------------------------------------------
       /cut1 |  -1.386294   .2877052                     -1.950186   -.8224026
       /cut2 |  -.6931472   .2604341                     -1.203589   -.1827058
       /cut3 |   1.098612   .2774642                      .5547925    1.642432
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation to odds ratios.

. disp "Intercepts: " log(1/4) ", " log(1/2) ", and " log(3)
Intercepts: -1.3862944, -.69314718, and 1.0986123
```

Regression of ordinal outcomes

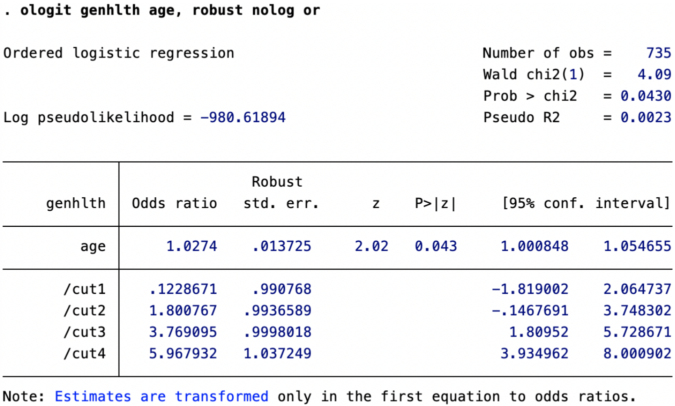
Example: General health in MRI study

- Let us use data from the MRI cohort to examine the association between age and participant's self-reported view of health.
  - X: age (years).
  - Y: view of own health
    - ⋆ $1 =$ excellent
    - ⋆ $2 =$ very good
    - ⋆ $3 =$ good
    - ⋆ $4 =$ fair
    - ⋆ $5 =$ poor
- Model (for $m = 1, \dots, 4$):

$$\text{logit}(\text{P}(Y \leq m | X = x)) = \beta_{0m} - \beta x.$$

Regression of ordinal outcomes

Stata: General health in MRI study



```
. ologit genhlth age, robust nolog or

Ordered logistic regression                             Number of obs =    735
                                                        Wald chi2(1)  =   4.09
                                                        Prob > chi2   = 0.0430
Log pseudolikelihood = -980.61894                       Pseudo R2     = 0.0023

------------------------------------------------------------------------------
             |               Robust
     genhlth | Odds ratio   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |     1.0274    .013725     2.02   0.043     1.000848    1.054655
-------------+----------------------------------------------------------------
       /cut1 |   .1228671    .990768                     -1.819002    2.064737
       /cut2 |   1.800767   .9936589                     -.1467691    3.748302
       /cut3 |   3.769095   .9998018                       1.80952    5.728671
       /cut4 |   5.967932   1.037249                      3.934962    8.000902
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation to odds ratios.
```

Regression of ordinal outcomes

Example: General health in MRI study

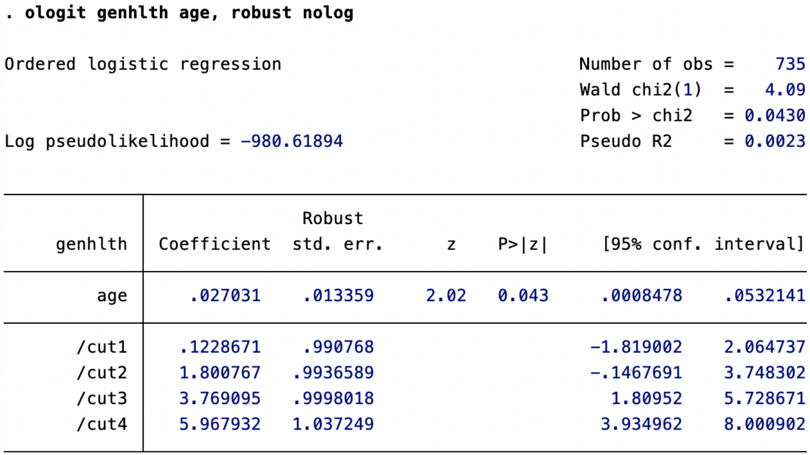
- Variables:
  - X: age (years).
  - Y: view of own health (1:5)
- Model (for $m = 1, \ldots, 4$):

$$\text{logit}(\text{P}(Y \leq m | X = x)) = \beta_{0m} - \beta x.$$

- As usual, let's get a few examples out of this one model:
  1. Estimate the odds of a good or better self-view of health among 85 year-olds.
  2. Estimate the proportion of 75 year-olds with a very good view their own health.
  3. Estimate the proportion of 70 year-olds with a fair or worse view their own health.
- To confirm our estimates of these values "by hand," it will be easier to have the untransformed output.

Regression of ordinal outcomes

Stata: General health in MRI study



```
. ologit genhlth age, robust nolog

Ordered logistic regression                         Number of obs =    735
                                                    Wald chi2(1)  =   4.09
                                                    Prob > chi2   = 0.0430
Log pseudolikelihood = -980.61894                   Pseudo R2     = 0.0023

------------------------------------------------------------------------------
             |               Robust
     genhlth | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |    .027031    .013359     2.02   0.043     .0008478    .0532141
-------------+----------------------------------------------------------------
       /cut1 |   .1228671    .990768                     -1.819002    2.064737
       /cut2 |   1.800767   .9936589                     -.1467691    3.748302
       /cut3 |   3.769095   .9998018                       1.80952    5.728671
       /cut4 |   5.967932   1.037249                      3.934962    8.000902
------------------------------------------------------------------------------
```

Regression of ordinal outcomes

Example: General health in MRI study

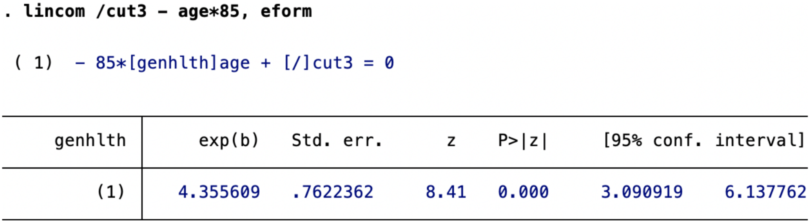
- Model: $\text{logit}(P(Y \leq m|X = x)) = \beta_{0m} - \beta x$.
- **Example 1**: Estimate the odds of a good or better self-view of health among 85 year-olds.
  - A good or better view means $M \leq 3$
  - By the model, $\log(O(Y \leq 3|X = 85)) = \beta_{03} - 85\beta$.
  - Therefore:

$$\begin{aligned} \widehat{O}(Y \leq 3|X = 85) &= \exp(\widehat{\beta}_{03} - 85\widehat{\beta}) \\ &= \exp(3.769095 - 85 \times 0.027031) \\ &= 4.3556. \end{aligned}$$

- We should be able to confirm this with the `lincom` command (which will also give us a confidence interval for this quantity).

Regression of ordinal outcomes

Stata: General health in MRI study



```
. lincom /cut3 - age*85, eform

 ( 1)  - 85*[genhlth]age + [/]cut3 = 0

------------------------------------------------------------------------------
     genhlth |     exp(b)   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         (1) |   4.355609   .7622362     8.41   0.000     3.090919    6.137762
------------------------------------------------------------------------------
```

Regression of ordinal outcomes

Example: General health in MRI study

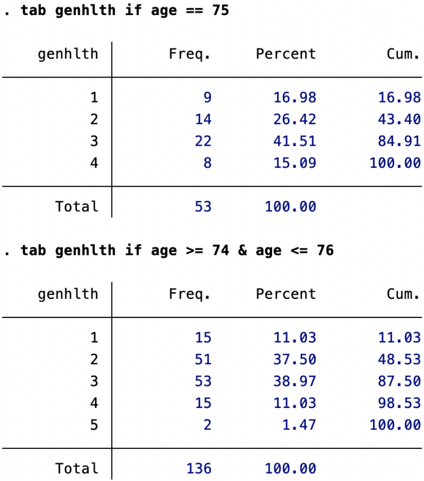
- Model: $\text{logit}(\text{P}(Y \leq m|X = x)) = \beta_{0m} - \beta x$.
- **Example 2**: Estimate the proportion of 75 year-olds with a very good view their own health.
  - A very good view means $M = 2$.
  - Note: $\text{P}(Y = 2|X = 75) = \text{P}(Y \leq 2|X = 75) - \text{P}(Y \leq 1|X = 75)$.
  - Model: $\text{P}(Y = 2|X = 75) = \text{expit}(\beta_{02} - 75\beta) - \text{expit}(\beta_{01} - 75\beta)$.
  - Therefore, $\widehat{\text{P}}(Y = 2|X = 75)$ can be expressed as:

$$\begin{aligned} & \text{expit}(\widehat{\beta}_{02} - 75\widehat{\beta}) - \text{expit}(\widehat{\beta}_{01} - 75\widehat{\beta}) \\ = \ & \text{expit}(1.800767 - 75 \times 0.027031) \\ & \quad - \text{expit}(0.1228671 - 75 \times 0.027031) \\ = \ & 0.44360 - 0.12960 = 0.314. \end{aligned}$$

- We can also see if this resembles the true proportion of patients with a very good view of their health at this age (or in a nearby range).

Regression of ordinal outcomes

Stata: General health in MRI study



```
. tab genhlth if age == 75

    genhlth |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |          9       16.98       16.98
          2 |         14       26.42       43.40
          3 |         22       41.51       84.91
          4 |          8       15.09      100.00
------------+-----------------------------------
      Total |         53      100.00

. tab genhlth if age >= 74 & age <= 76

    genhlth |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         15       11.03       11.03
          2 |         51       37.50       48.53
          3 |         53       38.97       87.50
          4 |         15       11.03       98.53
          5 |          2        1.47      100.00
------------+-----------------------------------
      Total |        136      100.00
```

Regression of ordinal outcomes

Example: General health in MRI study

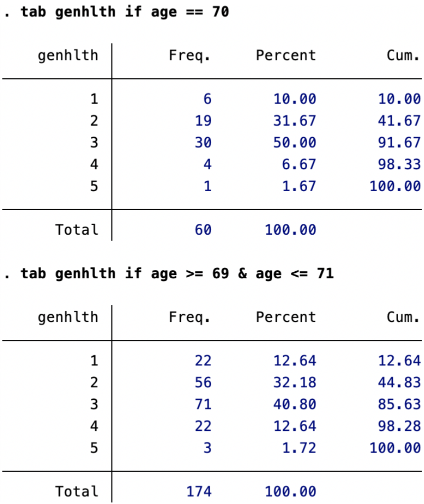
- Model: $\text{logit}(\text{P}(Y \leq m|X = x)) = \beta_{0m} - \beta x$.
- **Example 3**: Estimate the proportion of 70 year-olds with a fair or worse view their own health.
  - A fair or worse view means $M \geq 4$.
  - Note: $\text{P}(Y \geq 4|X = 70) = \text{P}(Y > 3|X = 70) = 1 - \text{P}(Y \leq 3|X = 70)$.
  - Model: $\text{P}(Y \geq 4|X = 70) = 1 - \text{expit}(\beta_{03} - 70\beta)$.
  - Therefore:

$$\begin{aligned}\widehat{\text{P}}(Y \geq 4|X = 70) &= 1 - \text{expit}(\widehat{\beta}_{03} - 70\widehat{\beta}) \\ &= 1 - \text{expit}(3.769095 - 70 \times 0.027031) = 0.133.\end{aligned}$$

- We can also see if this resembles the true proportion of patients with a fair or worse view of their health at this age (or in a nearby range).

Regression of ordinal outcomes

Stata: General health in MRI study



```
. tab genhlth if age == 70

    genhlth |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |          6       10.00       10.00
          2 |         19       31.67       41.67
          3 |         30       50.00       91.67
          4 |          4        6.67       98.33
          5 |          1        1.67      100.00
------------+-----------------------------------
      Total |         60      100.00

. tab genhlth if age >= 69 & age <= 71

    genhlth |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         22       12.64       12.64
          2 |         56       32.18       44.83
          3 |         71       40.80       85.63
          4 |         22       12.64       98.28
          5 |          3        1.72      100.00
------------+-----------------------------------
      Total |        174      100.00
```

Regression of ordinal outcomes

Additional thoughts:

- The cumulative logit imposes a proportional odds structure.
  - This goes a bit beyond the “parallel lines” assumption that we see in regression of continuous, binary, and multinomial outcomes.
  - The proportional odds assumption assumes that the odds ratios that compare subgroups differing in X by one unit are the same—not just regardless of X, but also regardless of the category of Y.
  - The baseline odds is an entire (discrete-valued) function that is estimated as a sequence of intercepts (one for each category of Y).
- Needless to say, you should be able to generalize the ideas of adjustment, interactions, categorical covariates, splines, transformations, etc. to models involving ordinal outcomes.

Table of Contents

1. Discrete outcomes

2. Regression of nominal outcomes

3. Regression of ordinal outcomes

4. Poisson regression (count outcomes)

Regression of count data

Ideas:

- Count examples: polyps in patient's colon during time between colonoscopies, pulmonary exacerbations experienced by a cycstic fibrosis patient during a year.
- Poisson distribution: # of events in a specified time (and space), parameterized by a constant rate, $\lambda > 0$.
  - $Y \sim \text{Poisson}(\lambda)$.

$$\text{P}(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!}; k = 0, 1, 2, \ldots$$

  - $\text{E}[Y] = \lambda$.
  - $\text{Var}[Y] = \lambda$.
- Note relationship between mean and variance.
- Most often summarize and compare response via *event rate*.

Regression of count data

Count data: Event rate

- Event rate: Expected # of events per unit of space-time.
- Must know interval of time, volume of space sampled.
- Often, we assume the counts follow a Poisson distribution, which can be derived from the following assumptions:
  - The expected number of events occurring in an interval of time is proportional to the size of the interval.
  - The probability that two events occur in an infinitesimally small interval of space-time is zero.
  - The number of events occurring in *separate* intervals of space-time are independent.
- Assumption of a constant rate with independence over separate intervals is often a very strong assumption.

Regression of count data

Regression of counts:

- When response variable represents counts of some event, we typically model the (log) rate using Poisson regression.
- Compares rates of response per space-time (e.g., person-years) across groups.
- Model: $\log(\mathrm{E}[Y|X = x]) = \beta_0 + \beta_1 x$.
  - $\exp(\beta_0)$: event rate among subgroup $X = 0$.
  - $\exp(\beta_1)$: ratio of event rates, comparing subgroups differing in X by one unit.
- We often choose to model the log-rate for the same reason we often choose to model the log-odds for binary outcomes: it makes the math work out nicer than other choices.

Regression of count data

Estimating equations: Poisson regression ($g(\boldsymbol{\mu}) = \log(\boldsymbol{\mu})$)

- To estimate $\boldsymbol{\beta}$, we solve the following equations for $\boldsymbol{\beta}$:

$$\mathbf{X}^T(\mathbf{y} - \exp(\mathbf{X}\boldsymbol{\beta})) = \mathbf{0}.$$

- This equation does not possess a closed-form solution. Iterative methods are used to estimate $\boldsymbol{\beta}$.

Regression of count data

Poisson regression:

- Model: $\log(\mathrm{E}[Y|X=x]) = \beta_0 + \beta_1 x$.
- The sandwich variance is robust to misspecification of the mean model and the mean-variance relationship.
  - Assuming a Poisson distribution for Y assumes that $\mathrm{E}[Y] = \mathrm{Var}[Y]$, though it could be that Y does not follow an exact Poisson distribution.
  - It could be, in general, that $\mathrm{E}[Y] = \phi \mathrm{Var}[Y]$ for some $\phi \neq 1$, a phenomenon often referred to as over- or under-dispersion.
  - There are methods to specifically *leverage* over- or under-dispersion (e.g., negative binomial regression), that we won't discuss in detail.
- Poisson regression in Stata: `poisson`.
  - Option `robust` provides sandwich variance
  - Option `irr` exponentiates to provide incidence rate ratios.

Regression of count data

Example: Chemotherapy (Background)

- Laboratory research of chemotherapy agents involves testing of the drugs in a culture of cells derived from a single cancer cell.
- A sample drawn from a liquid culture of some cell line is exposed to a new drug or combinations of new drugs at varying concentrations.
- Following an incubation, the resulting colonies of cells can be counted.
- Let's draw a simple example from the `chemo.csv` data set (which has a lot more nuances than what I am highlighting here) to evaluate doxorubicin as a possible chemotherapy agent.

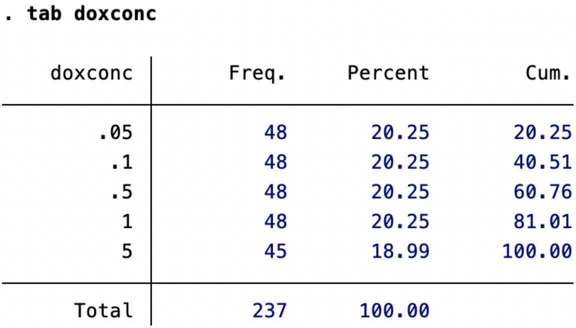
Regression of count data

Example: Chemotherapy

- Variables:
  - X: Concentration of doxorubicin (μmol/L) assigned to plate.
  - Y: Number of colonies following an incubation period.
- Only consider $X \geq 0.05$ for this example.
  - Concentrations of 0.05, 0.1, 0.5, 1, and 5 μmol/L.

Regression of count data

Example: Chemotherapy

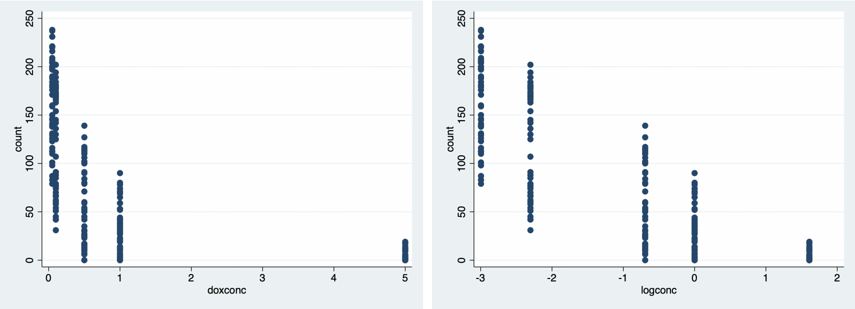


```
. tab doxconc

    doxconc |      Freq.     Percent        Cum.
------------+-----------------------------------
        .05 |         48       20.25       20.25
         .1 |         48       20.25       40.51
         .5 |         48       20.25       60.76
          1 |         48       20.25       81.01
          5 |         45       18.99      100.00
------------+-----------------------------------
      Total |        237      100.00
```

Regression of count data

Example: Chemotherapy

- `gen logconc = log(doxconc)`



Regression of count data

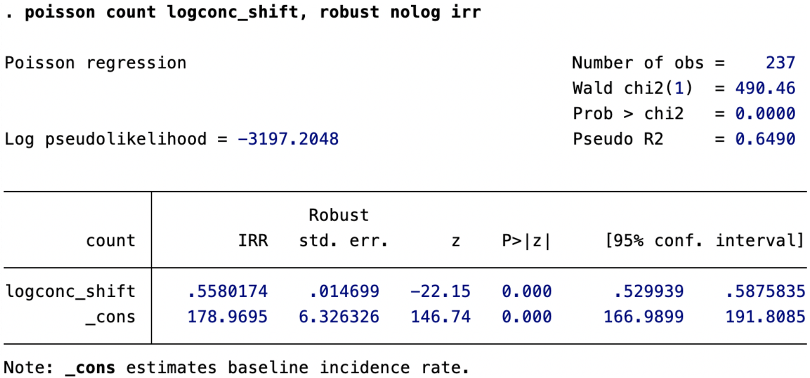
Example: Chemotherapy

- Variables:
  - X: Concentration of doxorubicin (μmol/L) assigned to plate.
    - Only consider $X \geq 0.05$ for this example.
  - Y: Number of remaining colonies following an incubation period.
- Model: $\log(\mathrm{E}[Y|X=x]) = \beta_0 + \beta_1(\log(x) - \log(0.05))$.
  - $\exp(\beta_0)$: The expected post-incubation colony frequency among plates assigned doxorubicin at a concentration of 0.05 μmol/L.
  - 2^{β_1}: Ratio of expected post-incubation colony frequency between plates differing in their doxorubicin concentration by a factor of two.

Regression of count data

Example: Chemotherapy

- `gen logconc_shift = log(doxconc) - log(0.05)`

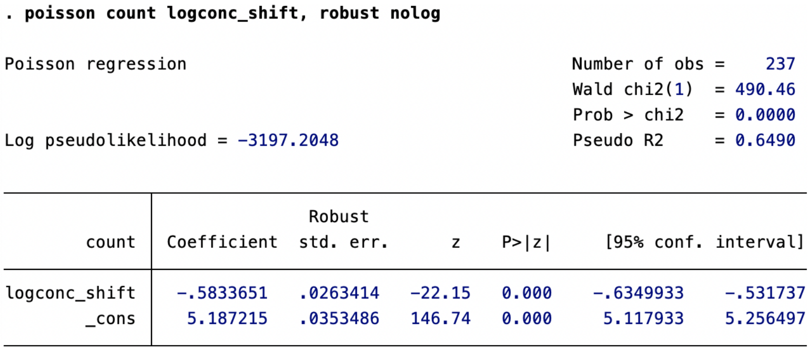


```
. poisson count logconc_shift, robust nolog irr

Poisson regression                                      Number of obs =    237
                                                        Wald chi2(1)  = 490.46
                                                        Prob > chi2   = 0.0000
Log pseudolikelihood = -3197.2048                       Pseudo R2     = 0.6490

------------------------------------------------------------------------------
             |               Robust
       count |        IRR   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
logconc_shift |   .5580174    .014699   -22.15   0.000      .529939    .5875835
       _cons |   178.9695   6.326326   146.74   0.000     166.9899    191.8085
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

- Because the concentration is log-transformed, the transformed output is not as directly helpful for characterizing the association.

Regression of count data

Example: Chemotherapy



```
. poisson count logconc_shift, robust nolog

Poisson regression                                      Number of obs =    237
                                                        Wald chi2(1)  = 490.46
                                                        Prob > chi2   = 0.0000
Log pseudolikelihood = -3197.2048                       Pseudo R2     = 0.6490

------------------------------------------------------------------------------
             |               Robust
       count | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
logconc_shift |  -.5833651   .0263414   -22.15   0.000    -.6349933    -.531737
       _cons |   5.187215   .0353486   146.74   0.000     5.117933    5.256497
------------------------------------------------------------------------------
```

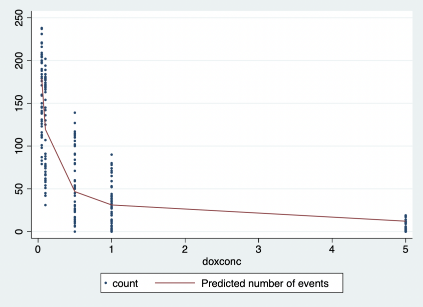
Regression of count data

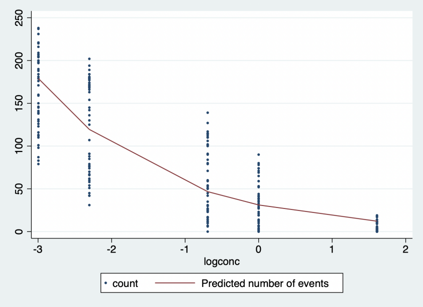
Example: Chemotherapy

- We can transform the coefficient estimate and the endpoints of its respective 95% confidence interval.
  - $\widehat{\beta}_1 = -0.5833651$ (95% CI: [-0.6349933, -0.531737]).
  - $2^{\widehat{\beta}_1} = 0.667$ (95% CI: [0.644, 0.692]).
- We estimate that each doubling of doxorubicin concentration is associated with a 33.3% reduction in expected post-incubation colony frequency (95% CI: [30.8%, 35.6%]; $p < 0.001$).

Regression of count data

Example: Chemotherapy





Regression of count data

Example: Chemotherapy

- It appears that over this range of concentrations, the Poisson model fits reasonably well.
  - In truth, the real data are better described by an S-shape per Michaelis-Menten mechanics.
- However, the question remains as to what the real advantage is of a Poisson model when I could have just computed the group-specific means, specifically given that the concentrations determined discretely in this example.

Regression of count data

Example: Chemotherapy

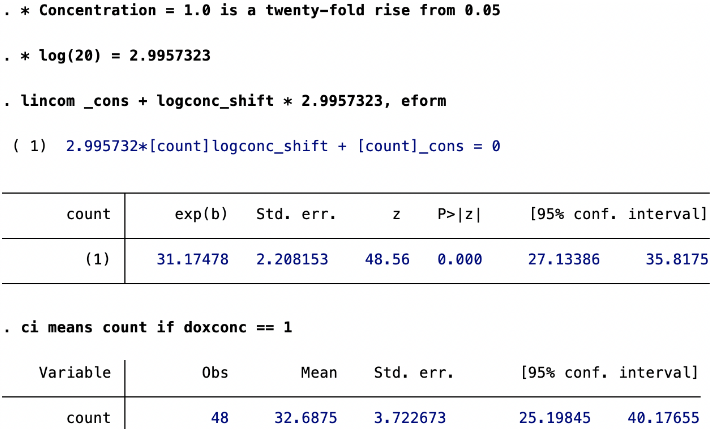
- Model-based estimate of expected frequency among a doxorubicin concentration of 1 μmol/L:

$$\begin{aligned}\log(\mathrm{E}[Y|X=1]) &= \beta_0 + \beta_1(\log(1) - \log(0.05)) \\ &= \beta_0 + 2.99573 \times \beta_1 \\ \Rightarrow \mathrm{E}[Y|X=1] &= \exp(\beta_0 + 2.99573 \times \beta_1).\end{aligned}$$

- The `lincom` command with the `eform` option will give us a point estimate and a 95% confidence interval for this quantity.
- Let's compare it to the simple (group-specific) estimate and 95% confidence interval.

Regression of count data

Example: Chemotherapy



```
. * Concentration = 1.0 is a twenty-fold rise from 0.05

. * log(20) = 2.9957323

. lincom _cons + logconc_shift * 2.9957323, eform

 ( 1)  2.995732*[count]logconc_shift + [count]_cons = 0

------------------------------------------------------------------------------
       count |     exp(b)   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         (1) |   31.17478   2.208153    48.56   0.000     27.13386     35.8175
------------------------------------------------------------------------------

. ci means count if doxconc == 1

    Variable |        Obs        Mean    Std. err.       [95% conf. interval]
-------------+---------------------------------------------------------------
       count |         48     32.6875    3.722673        25.19845    40.17655
```

- Modeling assumptions that are correct (or nearly so) are often rewarded with precision.

Offsets:

- The count outcomes are assumed to occur over a common range of space-time across observations.
  - This was satisfied in the chemotherapy example, which was on a plate-specific basis with a common incubation period.
- Suppose, as an example, that I seek to model the expected frequency of asthma exacerbations per year in children as a function of age, but that the amount of follow-up varies across children.
- All else being equal, a child with two years of follow-up will differ in their expected frequency as compared to a child with five years of follow-up.
- If you do not accommodate this sort of variability, the Poisson regression model will not be valid.
- Observation-specific regions of space-time can be accommodated by the inclusion of an *offset term*, which is essentially designed to level the playing field.

Regression of count data

Offsets:

- Consider a variable, W, that characterizes the observation-specific range of space-time.
- The Poisson model can be adjusted with a simple fix:

$$\log(\mathrm{E}[Y|X=x]/w) \quad = \quad \beta_0 + \beta_1 x.$$

- We can carry this forward just a bit further with basic properties of logarithms:

$$\begin{aligned} \log(\mathrm{E}[Y|X=x]) - \log(w) \quad &= \quad \beta_0 + \beta_1 x \\ \Rightarrow \log(\mathrm{E}[Y|X=x]) \quad &= \quad \beta_0 + 1 \times \log(w) + \beta_1 x. \end{aligned}$$

- Procedurally, this means we can accommodate variable space-time across observations by putting in a log-transformed W as a covariate in the model and force its corresponding coefficient to be one.

Regression of count data

Offsets:

- Model: $\log(\mathrm{E}[Y|X = x]) = \log(w) + \beta_0 + \beta_1 x$.
  - $\exp(\beta_0)$: expected frequency per one unit of W (event rate) among subgroup $X = 0$.
  - $\exp(\beta_1)$: ratio of event rates per one unit of W comparing subgroups differing in X by one unit.
- On your own: look up documentation for offset terms for Poisson regression models in Stata (be mindful of distinction between `offset` and `exposure`).

Regression of count data

Additional thoughts:

- Needless to say, you should be able to generalize the ideas of adjustment, interactions, categorical covariates, splines, transformations, etc. to models involving count outcomes.

Summary

Notes: Topics in this unit

- In this unit, we have learned all about regression of discrete outcomes.
  - Binary outcomes.
  - Nominal outcomes.
  - Ordinal outcomes.
  - Count outcomes.
- Much of the “regression math,” as we have been calling it, remains similar. Most discrete-outcome regression models involve some sort of transformation to the left-hand side, which means that we need to back-transform after the regression model is fit in order to obtain scientifically meaningful interpretations.
- Considerations regarding study design, assumptions, all need to be taken into account in order to make sensible choices.

Summary

Notes: Next unit

- Longitudinal data analysis!
  - Up until now, we have largely assumed that our observations were independent of one another.
  - What happens when you have, for instance, repeated measurements on the same individuals over time?