HUDM 5123 - Linear Models and Experimental Design
Regression Models for Count Data

1 Count Data

A variable is said to be measured on a "count" scale if it can take on a positive integer value in the range $\{0, 1, 2, \ldots\}$. Variables measured on the count scale are frequently encountered in the social and behavioral sciences. Some examples:

- Several national surveys ask high school students about past month frequency of risky behaviors. For example, "How many times in the past thirty days have you used marijuana?"
- A developmental psychologist runs a randomized experiment to test if having students practice structured debates in groups leads to better argumentative writing than a control group. All students write argumentative essays. One of the research hypotheses is that interventions group students will write more. As such, one of the outcomes is total number of idea units per essay.
- A researcher in the Department of Education has been tasked with reducing truancy in the public school system. To get an evidence-based sense of the factors at the school and student-levels that are important in predicting truancy, the researcher plans to pull about 4,000 student records to model the number of days absent per year (outcome) based on a number of predictors such as race, gender, socioeconomic status of student and school, student achievement, etc.

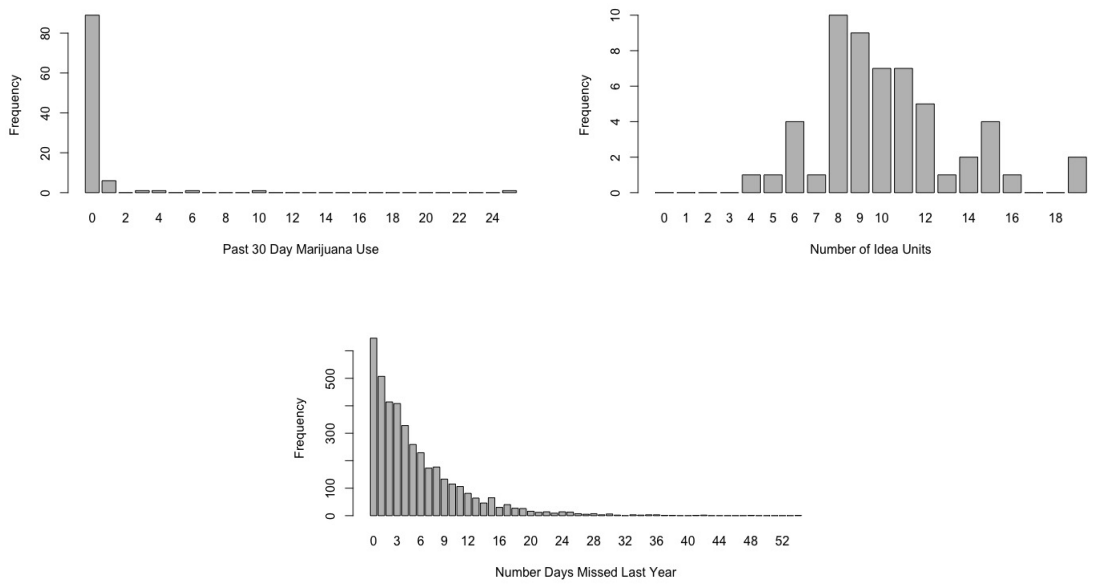

Figure 1: Count data examples

1.1 How Does OLS Regression Fail with Count Outcome Data?

If a count variable, like number of idea units above, is bounded away from zero, it may be reasonable to treat it as continuous and just use OLS. If you do, however, you may get predicted values that fall below zero. Furthermore, count variables, like 30 day marijuana use or days of school missed, that are centered near zero are typically heavily skewed to the right because they are bounded on the left at zero. OLS regression is inappropriate because the assumption of normally distributed residuals will be violated.

2 Generalized Linear Models (GLMs or GLiMs)

A generalized liner model (GLM) consists of three parts:

1. A random part, which specifies the conditional distribution of the outcome, Y_i, given the values of the predictors in the model.

2. A linear part, which is a linear function of the predictors,

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

3. An invertible *link function* g, which transforms the expected value of the response variable $\mu_i = E(Y_i)$, to the linear predictor:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

In the generalized linear modeling framework, there are often multiple ways to specify a link function, g, that transforms the linear part to the desired range. For binary outcome, if the link function is the log-odds, it leads to logistic regression. If the link function is the inverse of the cumulative normal distribution, it leads to probit regression, which is a similar approach. A link function is called "canonical" if has some particular statistical properties that make it easier to work with than other link functions. The logistic function is canonical for binomial outcome data.

For count outcome data, we need a function whose inverse transforms the range of the linear part, η_i, from $(-\infty, \infty)$ to the non-negative values $[0, \infty)$; that is, we seek g such that $g^{-1} : (-\infty, \infty) \to [0, \infty)$. The log function is a reasonable candidate that also turns out to be the canonical link for Poisson count outcome data.

Family	Canonical Link	Range of Y_i	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Gaussian	Identity	$(-\infty, \infty)$	μ_i	η_i
Binomial	Logit	$0, 1$	$\log\left[\frac{\mu_i}{1-\mu_i}\right]$	$\frac{\exp(\eta_i)}{1+\exp(\eta_i)}$
Poisson	Log	$0, 1, 2, \ldots$	$\log[\mu_i]$	$\exp(\eta_i)$

Table 1: Canonical link functions, inverses, and response ranges for three families

For a continuous and normally (Gaussian) distributed outcome variable, the canonical link is the identity function.

$$\begin{aligned} g(\mu_i) &= \eta_i \\ \mu_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}. \end{aligned}$$

For a binary outcome variable, the canonical link is the logit function.

$$\begin{aligned} g(\mu_i) &= \eta_i \\ \log\left[\frac{\mu_i}{1-\mu_i}\right] &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} \\ \frac{\mu_i}{1-\mu_i} &= \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}) \\ \mu_i &= \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}. \end{aligned}$$

For a count outcome variable, the canonical link is the log function.

$$\begin{aligned} g(\mu_i) &= \eta_i \\ \log(\mu_i) &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} \\ \mu_i &= \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}). \end{aligned}$$

3 Generalized Linear Models for Counts

To model data under the GLM, the distribution of the random part and the link function need to be specified. For count data, a basic starting point is to use the Poisson distribution with log link. The Poisson distribution is a discrete probability distribution based on a mean parameter $\mu > 0$. The mean and variance of the Poisson distribution are both equal to μ.

$$\Pr(Y = y) = \mu^y \left[\frac{\exp(-\mu)}{y!} \right], \text{ for } y = 0, 1, 2, \ldots$$

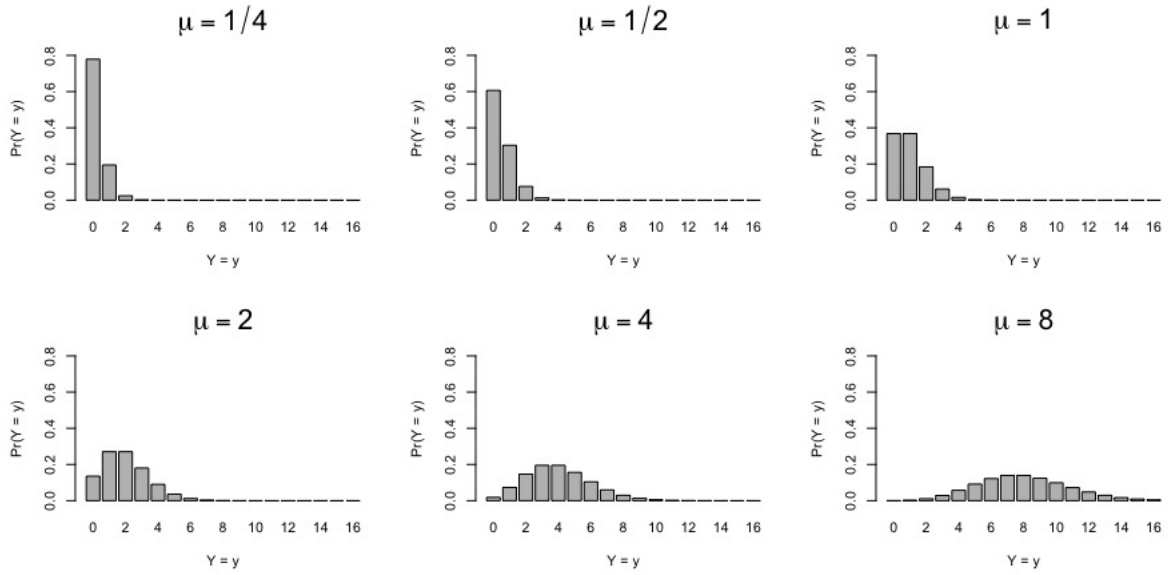

Figure 2: The Poisson distribution for several values of the μ parameter

The Poisson model, and other GLMs, are fit to data by the method of maximum likelihood, which provides both estimates of regression coefficients and asymptotic standard errors. Individual coefficients may be tested ($H_0 : \beta = 0$) by a Wald test by comparing $\beta/\text{SE}(\beta)$ to the standard normal distribution.

Overdispersion. The Poisson distribution has the rather unique property that its variance is identical to its mean. If the mean of a count outcome is close to its variance, great, you may be able to use Poisson regression. In practice we often encounter count data that have higher variance than mean. Data of this sort are called *overdispersed* and they require a more flexible distribution with an additional parameter to model the dispersion. The negative binomial distribution is one example of such a distribution.

Zero inflation. Beyond the issue of over-dispersion, another deviation from Poisson that is often seen in practice has to do with the frequency of zero values. In some cases where there are more zeros than can be explained by the Poisson model alone, we might hypothesize that there are actually two processes at play that can cause a response to be a zero. One is the usual Poisson process, which gives every unit some probability to be a zero. The other assumes that some units will respond with a zero no matter what. To put this in context, suppose we think of the past 30 day marijuana question. Suppose the

counts really do follow a Poisson process, but there are some respondents who, no matter how many times they actually used marijuana in the past 30 days, will respond "0" because they are afraid that they will get in trouble for admitting it, even though they are told the survey is anonymous. If we could predict who these respondents were based on other baseline covariates, we could model that process as distinct from the Poisson process.

Zero-inflated models accomplish this through two components: (1) a binary logistic model for membership in the unobserved (latent) class of always-zero responders and (2) a Poisson regression model for the not always-zero responders.

4 Real Data Example

A project developed and tested a web-based drug abuse prevention program for adolescent girls. The nationwide sample of 13- and 14-year- old girls (N = 788) was recruited via Facebook ads. Enrolled girls were randomly assigned to the intervention or control condition. All girls completed pretest measures online. Following pretest, intervention girls interacted with the 9-session, gender-specific prevention program online. The program aimed to reduce girlsâĂŹ drug use and associated risk factors by improving their cognitive and behavioral skills around such areas as coping with stress, managing mood, maintaining a healthy body image, and refusing drug use offers.

We will focus on the past 30 day cigarette usage. The data include a treatment factor (0 = control arm, 1 = treatment arm), age at enrollment, parent ed (0 = no parent has college degree, 1 = a parent does), grades (1 = mostly A's, 5 = mostly F's), and self-reported frequency counts of past 30 day cigarette usage at baseline and posttest.

We will begin by fitting ANOVA:

```
lm1 <- lm(THIRTYDAYCIG.2 ~ STUDY_ARM, data = dat)
Anova(lm1, type = 3)

Response: THIRTYDAYCIG.2
             Sum Sq  Df F value   Pr(>F)
(Intercept)   666.8   1 26.4749 3.41e-07 ***
STUDY_ARM     126.5   1  5.0222  0.02532 *
Residuals   18940.9 752
```

Parameter estimates:

```
summary(lm1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3127     0.2551   5.145 3.41e-07 ***
STUDY_ARM1   -0.8195     0.3657  -2.241   0.0253 *
```

We then fit ANCOVA, controlling for pretest (lost an observation due to missingness on pretest):

```
lm2 <- lm(THIRTYDAYCIG.2 ~ STUDY_ARM + THIRTYDAYCIG.1, data = dat)
Anova(lm2, type = 3)

Response: THIRTYDAYCIG.2
                Sum Sq  Df  F value    Pr(>F)
(Intercept)      167.0   1   9.4307  0.002211 **
STUDY_ARM         47.5   1   2.6829  0.101853
THIRTYDAYCIG.1  5599.3   1 316.1707 < 2.2e-16 ***
Residuals      13282.3 750
```

Parameter estimates:

```
summary(lm2)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.47656    0.15518   3.071  0.00221 **
STUDY_ARM1     -0.25164    0.15363  -1.638  0.10185
THIRTYDAYCIG.1  0.65636    0.03691  17.781  < 2e-16 ***
```

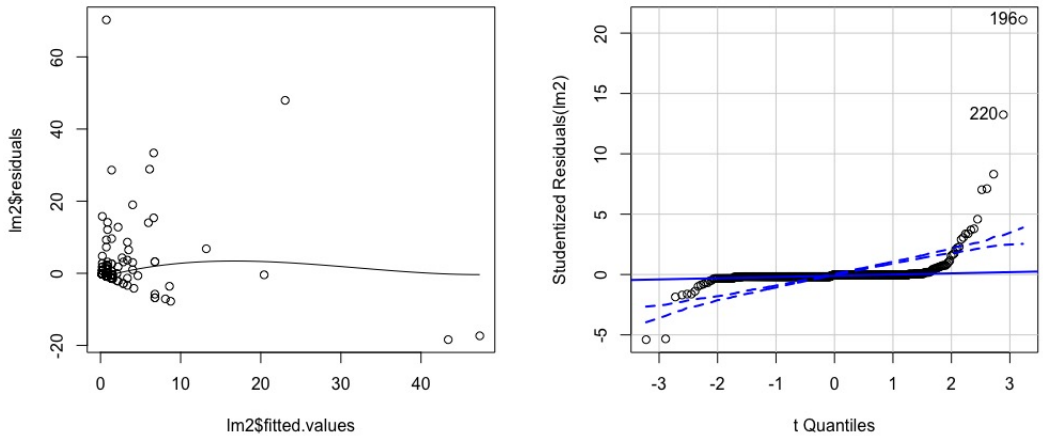

Figure 3: Plot of OLS residuals vs fitted values (left panel) and qq plot (right panel)

The qq plot is bad and suggests what we already knew: that the assumption of normally distributed residuals is not tenable due to the highly skewed count outcome. Let's try Poisson regression instead.

```
glm2 <- glm(THIRTYDAYCIG.2 ~ STUDY_ARM + THIRTYDAYCIG.1,
            family = "poisson",
            na.action = na.omit, data = dat)
summary(glm2)
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.405467   0.045640  -8.884  < 2e-16 ***
STUDY_ARM1     -0.332607   0.045349  -7.334 2.23e-13 ***
THIRTYDAYCIG.1  0.061146   0.001532  39.925  < 2e-16 ***
```

If the data are overdispersed, then Poisson is not the right stopping point for count modeling. A quick and dirty check is to examine the sample mean and variance of the outcome variable. This is not precise because the assumption is that the *conditional* mean and *conditional* variance, after controlling for predictors in the model, should be identical under the Poisson distribution assumption. However, it can nevertheless give us an idea, if only a crude one.

```
> mean(dat$THIRTYDAYCIG.2, na.rm = TRUE)
[1] 0.9137931
> var(dat$THIRTYDAYCIG.2, na.rm = TRUE)
[1] 25.32191
```

Here, the variance is about 25 times larger than the mean, which suggests the data are overdispersed. There is a statistical test for overdispersion that may be used in this scenario as well.

```
library(AER)
dispersiontest(glm1)

Overdispersion test

data:  glm1
z = 2.84, p-value = 0.002256
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  23.27658
```

Based on the output above, we conclude there is significant evidence of dispersion above and beyond that which would be expected under the Poisson model. Thus, let's use negative binomial modeling instead. The negative binomial regression model is based on the assumption that the errors follow a negative binomial distribution, which has a mean $E(Y_i) = \mu_i$ and $\text{Var}(Y_i) = \mu_i + \mu_i^2/\omega$, where ω is a scale parameter. Thus, the variance of Y grows larger as the mean increases at a rate that is faster than the mean.

```
nb1 <- glm.nb(THIRTYDAYCIG.2 ~ STUDY_ARM + THIRTYDAYCIG.1,
              control = glm.control(maxit = 1000),
              na.action = na.omit, data = dat)
summary(nb1)

Coefficients:
```

```
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.18693    0.15301  -7.757  8.7e-15 ***
STUDY_ARM1     -0.34800    0.15126  -2.301   0.0214 *
THIRTYDAYCIG.1  0.45852    0.03286  13.952  < 2e-16 ***
```

Finally, zero-inflation may also be an issue with these data. This is an obvious concern given the preponderance of zeros.

```
fit.ZINB <- zeroinfl(THIRTYDAYCIG.2 ~ STUDY_ARM + THIRTYDAYCIG.1 |
                     THIRTYDAYCIG.1 + AGE_YRS.1 + EDUC2.1 + GRADES.1,
                     dist = "negbin", na.action = na.omit, data = dat)

Count model coefficients (negbin with log link):
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.81085    0.19309   4.199 2.68e-05 ***
STUDY_ARM1     -0.32480    0.15019  -2.163  0.03057 *
THIRTYDAYCIG.1  0.08819    0.02966   2.973  0.00295 **
Log(theta)     -1.06487    0.16843  -6.322 2.57e-10 ***

Zero-inflation model coefficients (binomial with logit link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.02923    4.03778  -0.007   0.9942
THIRTYDAYCIG.1 -13.78789  180.15123  -0.077   0.9390
AGE_YRS.1        0.22404    0.29528   0.759   0.4480
EDUC2.1         -0.05619    0.39730  -0.141   0.8875
GRADES.1        -0.50191    0.22126  -2.268   0.0233 *
```

Fit indices such as the AIC and BIC are useful for comparing models that are not nested (like Poisson to ZINB, for example). AIC is short for Akaike information criterion. BIC is short for Bayesian information criterion. Both are used to measure overall model parsimony for models fit to the same data.

$$\text{AIC} = -2logL + 2p \qquad \text{and} \qquad \text{BIC} = -2logL + log(n)p,$$

where $logL$ is the value of the log of the likelihood at the maximum likelihood estimator for the given model, p is the number of parameters, and n is the sample size. Here are the results for AIC and BIC for the four analyses we ran (OLS, Poisson, negative binomial, zero-inflated negative binomial:

```
> c(AIC(lm2), AIC(glm2), AIC(nb2), aic(zinb2))
[1] 4306.1257 3403.9118 1007.2926  826.2348
> c(BIC(lm2), BIC(glm2), BIC(nb2), bic(zinb2))
[1] 4324.622 3417.784 1025.789  867.130
```

In this case the AIC and BIC agree that the zero-inflated model is superior (lower is better).