HUDM 5123 - Linear Models and Experimental Design
08 - Categorical Outcome

1 Linear Regression and Categorical Outcomes

Up to this point, all the models we have studied have been based on fitting and comparing multiple linear regression models of the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i,$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2_\epsilon)$. The assumptions encoded in this statement about the residuals are

1. linearity (i.e., $E[\epsilon|X] = 0$),
2. constant variance (i.e., $\text{var}[\epsilon|X] = \sigma^2_\epsilon$),
3. normality of residuals, and
4. mutual independence of residuals.

One of the consequences of this set of assumptions is that the outcome variable is assumed to be measured on a continuous scale. What if the outcome were dichotomous (i.e., a two-category variable with categories such as voted or did not vote) or unordered polytomous (i.e., a multicategory variable with categories such as democratic, independent, republican) instead? Why do we need new statistical machinery to model dichotomous data? What happens if we use linear regression modeling with a dichotomous outcome?

Let $\mathbf{X} = X_1, X_2, \ldots, X_p$ represent the matrix of covariate values and let $\mathbf{X}_i = X_{i1}, X_{i2}, \ldots, X_{ip}$ represent the covariate vector for unit i. The predicted values for the linear probability model are

$$\hat{Y}_i = E(Y_i|\mathbf{X}_i = \mathbf{x}_i) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip}.$$

1.1 Example with Acupuncture Data

As a reminder, the data for today's notes come from a randomized experiment to study the efficacy of acupuncture for treating headaches. Results of the trial were published in the British Medical Journal in 2004. You may view the paper at the following link: `http://www.bmj.com/content/328/7442/744.full`. The data set includes 301 cases, 140 control (no acupuncture) and 161 treated (acupuncture). Participants were randomly assigned to groups. Variable names and descriptions are as follows:

- `age`; age in years
- `sex`; male = 0, female = 1
- `migraine`; diagnosis of migraines = 1, diagnosis of tension-type headaches = 0
- `chronicity`; number of years of headache disorder at baseline

- `acupuncturist`; ID for acupuncture provider
- `group`; acupuncture treatment group = 1, control group = 0
- `pk1`; headache severity rating at baseline
- `pk5`; headache severity rating 1 year later

There is one additional variable (that I made up for this lecture), called `remission`, that was coded as 1 if the participant's headache disorder was considered to be in remission at the end of the study, and 0 if not. The two-way table of group assignment by remission status:

Table 1: Two-way table of treatment by remission status

	No Acupuncture	Acupuncture	Total
No Remission	85	66	151
Remission	55	95	150
Total	140	161	301

Boxplots of baseline headache severity by group assignment and remission status after one year:

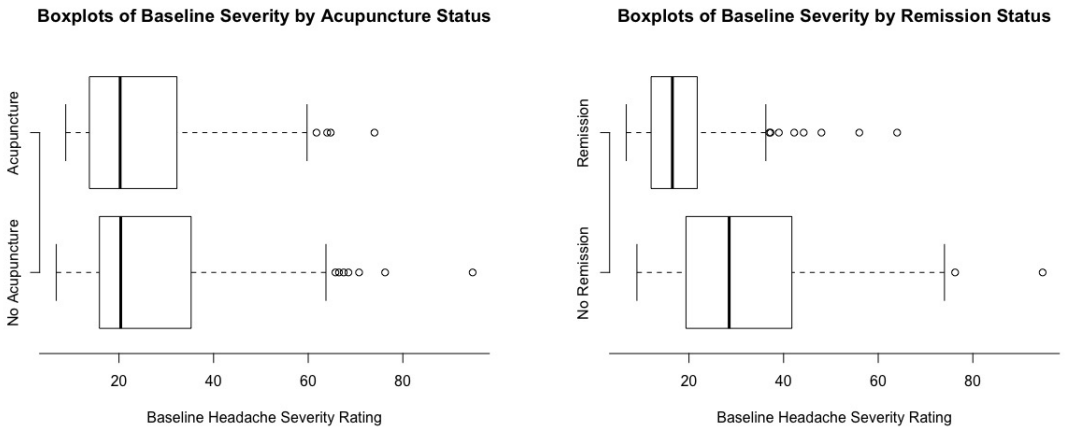

Figure 1: Parallel boxplots of baseline headache severity by group assignment (left panel) and baseline headache severity by remission status after one year (right panel); data from both acupuncture groups are included here

What is the relationship between baseline headache severity (predictor) and remission status (outcome), if any? We will investigate using multiple linear regression, even though the outcome is dichotomous. The scatterplot below is similar to the boxplot except that the scatterplot shows the actual data instead of five-number summaries. The plot also includes the simple linear regression line of best fit.
Recall that the best fit (i.e., prediction) equation is determined by taking the conditional expectation of both sides of the regression model, given the observed covariates.

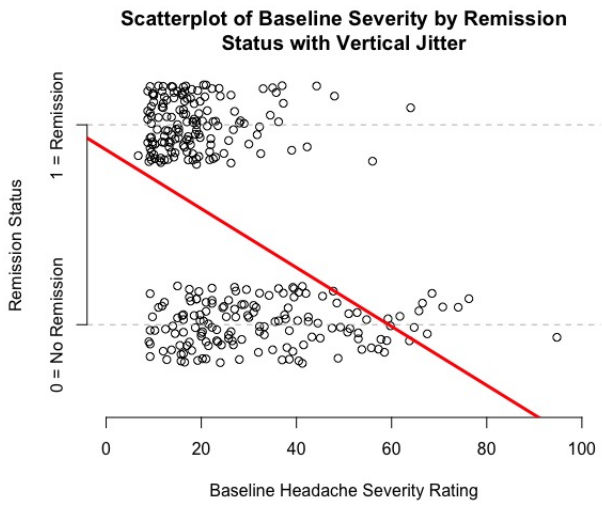

Figure 2: Scatterplot of remission status on baseline headache severity rating with regression line; points have been vertically jittered for better visibility

$$\hat{Y}_i = E(\text{Remission}_i|\text{PK1}_i = \text{pk1}_i) = \hat{\beta}_0 + \hat{\beta}_1\text{PK1}_i.$$

Also, recall that remission is a dichotomous variable. What does it mean to take the average of a dichotomous variable? Consider the sample average based on 20 observations where 12 were 1s (remission) and 8 were 0s (no remission). The sample average is $12/20 = 0.6$. In fact, by taking the average (expected value) of a dichotomous variable, you get the probability that the variable takes on the value 1. So, for the prediction equation above, we get the following:

$$\begin{aligned} E(\text{Remission}_i|\text{PK1}_i = \text{pk1}_i) &= \hat{\beta}_0 + \hat{\beta}_1\text{PK1}_i \\ p(\text{Remission}_i = 1|\text{PK1}_i = \text{pk1}_i) &= \hat{\beta}_0 + \hat{\beta}_1\text{PK1}_i \end{aligned}$$

That is, the conditional expectation of the remission variable given baseline severity may be interpreted as the conditional probability of remission given baseline severity. That is, for a dichotomous outcome, the predicted values are *conditional probabilities*. For example, according to our model fit, the probability of remission for a participant with a baseline score of 25 is about .5 (i.e., 50% chance of remission). Here are some other conditional probabilities predicted by the model given a certain baseline severity rating:

Note that we end up with negative predicted values. That is, the model predicts that a person with baseline headache severity of 70 has a -.15 probability of remission. Unfortunately, negative probabilities are not interpretable; this is a problem. The other major issue with using linear regression for dichotomous outcome variables has to do with the residuals.

Note that for any value of the predictor variable, `pk1`, there are only two possible values for the residuals. If the predicted value at `pk1` is $\hat{y}_1$, then the residual values for that point

Table 2: Table of model-based conditional probabilities of remission given baseline severity values

Baseline Severity	Conditional Probability
0	0.88
10	0.73
20	0.58
30	0.43
40	0.29
50	0.14
60	-0.01
70	-0.15
80	-0.30
90	-0.45
100	-0.60

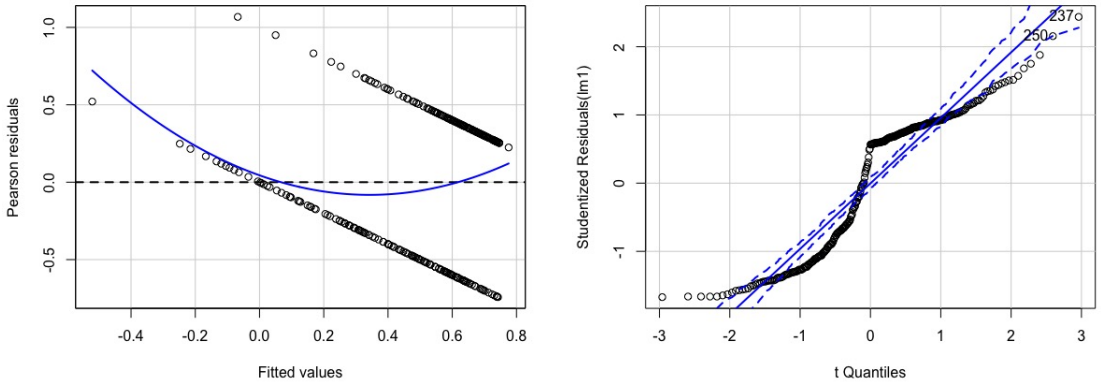

Figure 3: Studentized residuals vs fitted values (left panel) and qqplot (right panel)

are either 1 - $\hat{y}_1$ or 0 - $\hat{y}_1$. For example, for `pk1` = 20, the predicted value is .58. There are only two actual values that participants could have reported: in remission (1) or not (0), so the residuals will necessarily either be 1 - .58 = .42 or 0 - .58 = -.58. This leads to two obvious violations of the assumptions. First, the residuals cannot be normally distributed because of the dichotomous values; second, they will not have constant variance across the values of `pk1`.

Finally, interpretation of slope coefficients can be a challenge for linear models fit to a dichotomous outcome. For our example, the fitted values yield the following prediction equation:

$$p(\text{Remission}_i = 1|\text{PK1}_i = \text{pk1}_i) = 0.875 - 0.015\text{pk1}_i,$$

which implies that a 1 unit increase in baseline severity is associated with a 1.5% decrease in probability of remission. While this seems sensible for values of `pk1` in the middle of the domain, it becomes nonsensical once the predicted probabilities are impossible.

In summary, we have identified two major challenges to using linear regression methods with dichotomous outcome data.

1. The conditional expectation of a dichotomous variable is a probability, and probabilities are bound between 0 and 1, inclusive. A linear regression fit will result in *impossible predicted values* including negatives and values greater than 1.

2. Assumptions required for the validity of inferences made with linear regression are demonstrably false when the outcome is dichotomous. In particular, the constant variance and normality assumptions are necessarily violated when working with a dichotomous dependent variable.

Nevertheless, some authors have argued that it is still acceptable to use linear regression to model a dichotomous outcome with some important caveats. Lumley, Diehr, Emerson, & Chen (2002; *Annual Review of Public Health*) `https://www.ncbi.nlm.nih.gov/pubmed/11910059` note the following on p. 162:

> When the dependent variable is binary, the most common analytic method is logistic regression. In this approach the assumptions fit the data. Further, the (exponentials of the) regression parameters can be interpreted as odds ratios, which are nearly identical to relative risks when the event under study is rare.
>
> Another possible approach is least-squares linear regression, letting Y be the 0/1 binary variable. Such an approach is not usually considered appropriate because Y is not Normally distributed; however, the Central Limit Theorem ensures that the regression coefficients will be Normally distributed for large enough samples. Regression estimates would be a weighted sum of the Y's, which are 0's and 1's. The usual rule for the binomial distribution is that proportions are approximately Normal if np > 5 and n(1 - p) > 5, which should hold for the large data sets we are considering. Another objection to the linear regression approach is that estimated proportions can be below 0 or greater than 1. This is a problem if the goal is to predict a probability for an individual, and the sample is small. It will rarely be a problem when the goal is to assess the effects of independent variables on the outcome. A final objection is that the

homoscedasticity assumption is violated, since the variance is a function of the mean. The usual rule of thumb is that if proportions are between, say, 0.2 and 0.8, the variance is approximately constant and heteroscedasticity is not a serious problem.

So far we have only looked at the relationship between baseline severity and remission; that is, we have ignored the treatment variable. The linear fit including the treatment variable:

$$p(\text{Remission}_i = 1|\text{PK1}_i = \text{pk1}_i) = 0.776 - 0.014\text{pk1}_i + .167\text{group}_i.$$

The model suggests that those who received acupuncture treatment were about 17% more likely to experience remission one year later than those who did not (t = 3.27; p = .0012). The intercept suggests that the probability of remission for a person in the control group with baseline severity score of zero is about 78%. While these interpretations are sensible, we need to note the caveat that they only apply for participants in the middle range of the scores for `pk1`. For participants with more extreme baseline headache severity scores, these results will not apply, and, furthermore, predicted values will be nonsensical.

1.2 Multicategory Outcomes

So far we have only dealt with the case of dichotomous dependent variable. Suppose the dependent variable has three categories, none of which has numerical meaning (i.e., they are truly nominal/categorical) such as democrat, republican, and independent. Dummy coding would create a multivariate (bivariate in this case) outcome variable. While it might be possible to use linear regression to regress each category on the other to study relationships, this approach would only use subsets of data and would run into the same problems discussed above with dichotomous dependent variable. Next class we will discuss logistic regression and extensions. As we will see, logistic regression is built around the *logistic transformation*, which puts the probabilities on the right scale and makes nonsensical predicted probabilities impossible.

2 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.

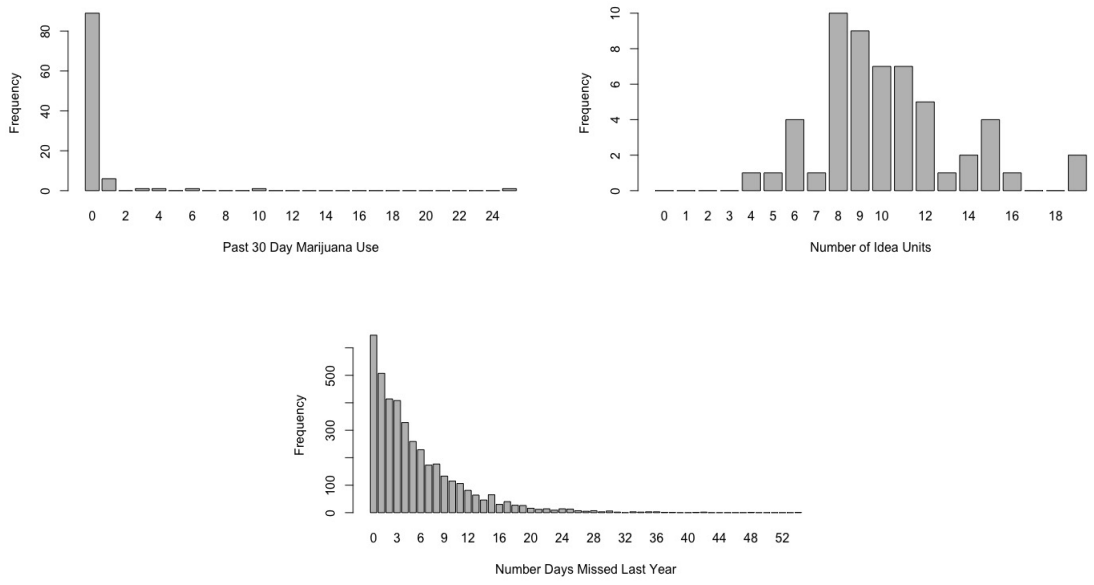

Figure 4: Count data examples

2.1 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 use linear regression. 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 assumptions of normally distributed residuals. It is also often the case that count-type dependent variables have variance that depends on the mean such that, for example, for lower counts the data are less variable but for higher counts they are more variable. Clearly these types of dependencies violate the constant variance assumption.

2.2 Example with Healthcare Data

Data come from a paper by Deb and Travedi (1997; *Journal of Applied Econometrics*) `http://www.econ.queensu.ca/jae/1997-v12.3/deb-trivedi/` include 4406 participants of age 66 or older who are covered by the public insurance program Medicare in the US. Data were prepared and described by Zeiles, Kleiber, & Jackman `https://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf`. From their paper (p. 9):

> The objective is to model the demand for medical care - as captured by the number of physician/non-physician office and hospital outpatient visits - by the covariates available for the patients. Here, we adopt the number of physician office visits `ofp` as the dependent variable and use the health status variables `hosp` (number of hospital stays), `health` (self-perceived health status), `numchron` (number of chronic conditions), as well as the socioeconomic variables `gender`, `school` (number of years of education), and `privins` (private insurance indicator) as regressors.

A barplot of the outcome variable, number of physician office visits, reveals heavy right skew and a maximum value of 89:

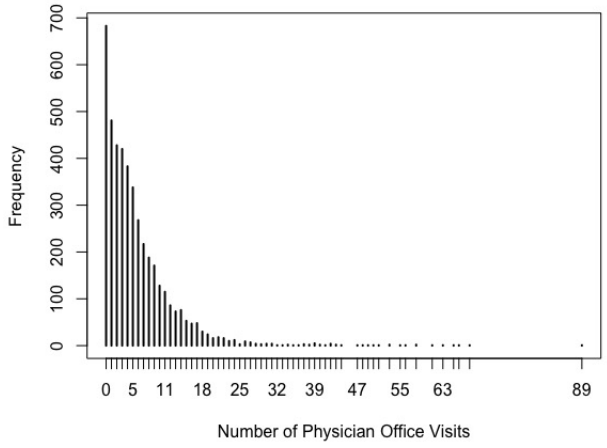

Figure 5: Barplot of number of physician office visits

With skewed outcome data a log transformation is often recommended. Here, there are

0 values, and log(0) is undefined. Remember, the log with base b is defined as follows:

$$\log_b(x) = a \iff x = b^a.$$

As an example,

$$\log_2(8) = 3 \text{ because } 2^3 = 8.$$

The log of 0 is undefined because it results in the following:

$$\log_2(0) = a \iff 0 = 2^a,$$

but there is no power, a, that will make 2^a be equal to 0. Thus, to take the log of the outcome, we need to shift the zero values over to the right so they are positive. One approach is to simply add 1/2 to the data before taking the log. Here is a plot of the bivariate relationship between the log of 1/2 plus the outcome and the predictor `numchron`:

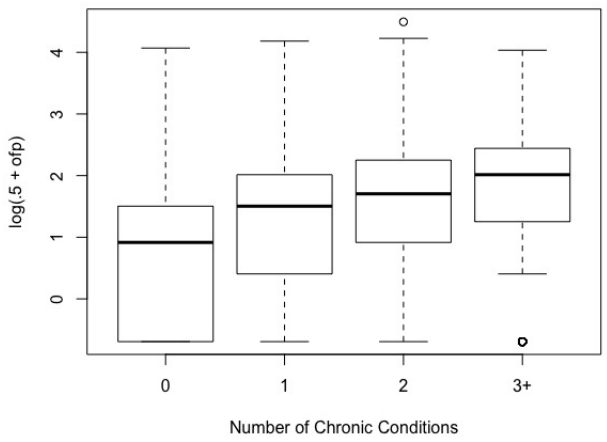

Figure 6: Boxplots of log of 1/2 plus number of office visits on number of chronic conditions categories

Dummy-coding the `numchron` variable with 0 held out as the reference category yields the following prediction equation:

$$\log(\text{ofp}_i + 0.5) = 0.76 + 0.47D_{1i} + 0.80D_{2i} + 1.04D_{3+i}.$$

Residual plots reveal normality and constant variance assumptions are violated:

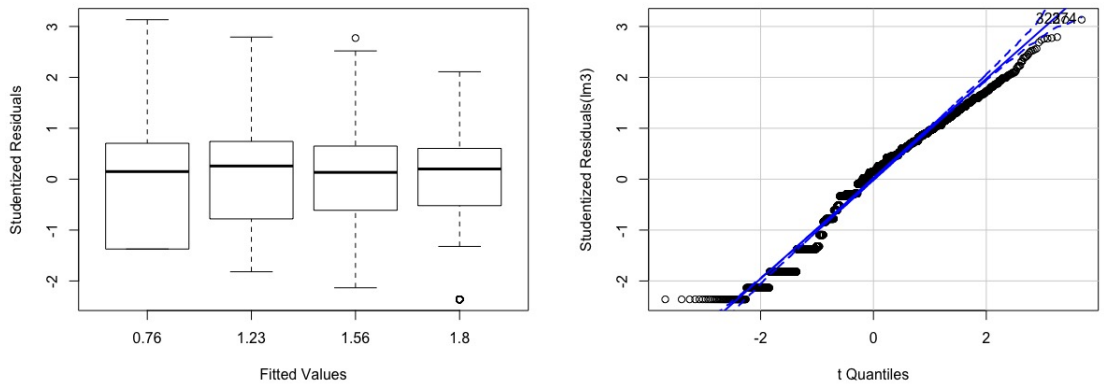

Figure 7: Residual plots for count data fit with linear model after log transformation of outcome

3 Lab

For lab today, you will replicate the plots and analysis run above with ECLS data. The variable `PROF5` is a dichotomous variable representing math proficiency at grade 5, where proficient = 1 and not proficient = 0, the variable `SPED` represents exposure to special education (1) or not (0), and variable `MATHK` is the kindergarten mathematics baseline proficiency score.

- Download and load into R the data file called 'ecls2.Rdata' and examine the data.
- Create horizontal (`horizontal = TRUE`) boxplots of baseline math proficiency by (a) special education exposure status and (b) 5th grade math proficiency.
- Create a scatterplot of 5th grade proficiency on baseline score with a vertical jitter.
- Estimate the linear regression and add the prediction line to the plot in red.
- Interpret results from the regression.
- Add special education status as a predictor and interpret results of all three model coefficients in context.