Applied Quantitative Methods for the Social Sciences II
MA in Social Sciences, UC3M, Spring 2026

Assignment 9: Other Outcomes (Ordinal, Multinomial, Count, Survival)

Instructions:

- Deadline: **April 16, before class**
- Submit your work in a separate folder in your GitHub repository
  - You can include only the R file or additional ones (e.g. pdf with results)
- **Always use comments** in your R code – and use them to answer questions
- You are encouraged to work together, but each person must submit their own code
- Plan is to start Part 1 in class and complete Part 2 at home
- I'll upload a solution file to the website after next class

Contents

1 Part 1: In-Class (Ordinal, Multinomial, and Count Outcomes)

In this lab we use the British Electoral Panel Study (BEPS), a survey of British voters conducted around the 1997 general election. We model two outcomes: (1) voters' perceptions of national economic conditions (an ordinal 1–5 scale), and (2) their vote choice (a nominal three-category outcome). Both require models that go beyond binary logistic regression. Recall from class that the key question before choosing a model is: *what is the data-generating process for your outcome?* Ordinal outcomes have a natural ordering but not equal spacing; nominal outcomes have unordered categories.

Load the required packages and data with:

```
library(carData)
library(MASS)
library(nnet)
library(marginaleffects)
data(BEPS)
```

Key variables in `BEPS` used in this assignment:

- `vote` — vote choice (factor: "Conservative", "Labour", "Liberal Democrat")
- `economic.cond.national` — national economic conditions (integer 1–5: 1 = got much worse, 5 = got much better)
- `economic.cond.household` — household economic conditions (integer 1–5)
- `Blair` — feelings about Tony Blair (integer 1–5, higher = more positive)
- `Hague` — feelings about William Hague (integer 1–5)
- `Kennedy` — feelings about Charles Kennedy (integer 1–5)
- `Europe` — attitude toward European integration (integer 1–11: 1 = strongly opposed, 11 = strongly in favor)
- `political.knowledge` — political knowledge (integer 0–3)
- `gender` — gender (factor: "female", "male")
- `age` — age in years

1.1 Ordered logit: perceptions of the national economy

We first model respondents' perceptions of national economic conditions as an ordinal outcome. Recall from class: OLS treats categories as equally spaced on the underlying scale, which is almost certainly wrong for survey Likert items. Ordered logit instead estimates both regression coefficients and threshold parameters ($\tau_1, \tau_2, \ldots$) that cut the latent continuous propensity Y^* into the observed ordered categories.

a) Explore the `economic.cond.national` variable and convert it to an ordered factor:

```
table(BEPS$economic.cond.national)
BEPS$econ_ord = factor(BEPS$economic.cond.national, ordered = TRUE)
```

In a comment, report the distribution across the five categories. Which category is most common? In a second comment, explain why using OLS on this variable would be problematic: think about the equal-spacing assumption and what it implies about the difference between "got much worse" (1) and "got a little worse" (2) versus the difference between "stayed the same" (3) and "got a little better" (4).

b) Fit an ordered logit model predicting `econ_ord` from `age`, `gender`, `Europe`, and `political.knowledg`

```
m_ologit = polr(econ_ord ~ age + gender + Europe + political.knowledge,
                data = BEPS, Hess = TRUE)
summary(m_ologit)
```

In a comment, report the raw coefficient on `Europe` and its sign. Based on the sign convention, does higher support for European integration predict more optimistic views of the national economy?

c) Compute average marginal effects (AMEs) using `marginaleffects`:

```
avg_slopes(m_ologit)
```

This returns one AME per predictor per response category. In a comment, interpret the AME for `Europe` on the probability of each category. Does a one-unit increase in pro-Europe attitude increase or decrease the probability of perceiving the economy as improved (category 4 or 5)? By approximately how much on average across respondents? Note that the AMEs across all five categories for any given predictor must sum to zero — use this as a sanity check.

d) Compute predicted probabilities for the five response categories at the mean of all covariates, separately for male and female respondents:

```
predictions(m_ologit, newdata = datagrid(gender = c("female", "male")))
```

In a comment, compare the predicted probabilities for the most pessimistic category (1 = got much worse) and the most optimistic category (5 = got much better) for each gender. Are there notable differences by gender? What does this suggest about gender gaps in economic perceptions?

1.2 Multinomial logit: vote choice

We now turn to vote choice (Conservative, Labour, Liberal Democrat) — a nominal three-category outcome with no natural ordering. Recall from class that the multinomial logit estimates $J - 1$ sets of coefficients, each comparing one category to a reference category. All predicted probabilities sum to 1, and the model does *not* impose any ordering on the alternatives.

a) Set Conservative as the reference category and fit a multinomial logit predicting `vote` from economic assessments and leader evaluations:

```
BEPS$vote = relevel(BEPS$vote, ref = "Conservative")
m_mlogit = multinom(vote ~ economic.cond.national + Blair + Hague +
                           Kennedy + Europe, data = BEPS, trace = FALSE)
summary(m_mlogit)
```

The model produces two sets of coefficients: Labour vs. Conservative and Liberal Democrat vs. Conservative. In a comment, describe the direction of the `Blair` coefficient in the Labour vs. Conservative equation. What does a positive coefficient on `Blair` (feelings toward Tony Blair) in the Labour equation imply about the relationship between Blair approval and the likelihood of voting Labour relative to Conservative?

b) Compute AMEs across all predictors and all outcome categories:

```
avg_slopes(m_mlogit)
```

In a comment, report the AME of `Blair` on the probability of voting Labour. Interpret it in plain language: holding other variables constant, how does a one-unit increase in Blair approval change the probability of voting Labour on average across respondents?

c) The multinomial logit assumes **Independence of Irrelevant Alternatives (IIA)**: the odds ratio between any two alternatives is unaffected by the presence or absence of other alternatives. Recall from class the red bus / blue bus example, where IIA fails because two alternatives (red bus and blue bus) are close substitutes. In a comment of 2–3 sentences, explain what IIA means for this application with Conservative, Labour, and Liberal Democrat as alternatives. Do you think IIA is likely to hold here — or are any two of these parties close substitutes in the minds of British voters? Explain your reasoning.

We now analyze the number of articles published by biochemistry PhD students in the last three years of their doctorate, using the `bioChemists` dataset from the `pscl` package. The outcome (`art`) is a non-negative integer count. Recall from class that count outcomes have a natural lower bound of zero and cannot take negative values, which makes OLS inappropriate. The natural starting model is Poisson regression; we then diagnose and address overdispersion using the negative binomial.

Load the required packages and data with:

```
library(pscl)
library(AER)
library(MASS)
library(marginaleffects)
data(bioChemists)
```

Key variables in `bioChemists`:

- `art` — articles published in last 3 years of PhD (integer, count outcome)
- `fem` — gender (factor: "Men", "Women")

- `mar` — marital status (factor: "Single", "Married")
- `kid5` — number of children under age 5 (integer 0–3)
- `phd` — prestige of PhD program (numeric)
- `ment` — articles by mentor in last 3 years (integer, count)

1.3 Poisson regression: publication counts

a) Explore the outcome variable `art`:

```
summary(bioChemists$art)
var(bioChemists$art)

pdf("art_histogram.pdf", width = 6, height = 4)
hist(bioChemists$art, breaks = 20, main = "Distribution of articles",
     xlab = "Number of articles", col = "gray80")
dev.off()
```

In a comment, report the mean and variance of `art`. A key diagnostic for count data is whether the variance substantially exceeds the mean — this is called **overdispersion** and violates the Poisson assumption that mean equals variance. Note whether you observe this pattern here.

b) Fit a Poisson regression of `art` on all predictors:

```
m_pois = glm(art ~ fem + mar + kid5 + phd + ment,
             data = bioChemists, family = poisson)
summary(m_pois)
```

In a comment, answer the following two questions: (1) Report the coefficient on `ment` and exponentiate it with `exp()` to obtain the **incidence rate ratio** (IRR). Interpret it: a one-unit increase in mentor articles multiplies expected student articles by approximately how much? (2) Report the residual deviance and degrees of freedom from the `summary()` output and compute their ratio. Recall from class that under a correctly specified Poisson model this ratio should be close to 1; a ratio substantially above 1 (say, > 2) suggests overdispersion.

c) Test for overdispersion formally:

```
dispersiontest(m_pois)
```

In a comment, report the estimated dispersion parameter and the p-value. Is there statistically significant evidence of overdispersion? What does this imply for the validity of the Poisson standard errors you computed above?

1.4 Negative binomial regression

The negative binomial (NB) model generalizes Poisson by adding a dispersion parameter θ that allows the variance to exceed the mean: $\text{Var}(Y_i) = \mu_i + \mu_i^2/\theta$. When $\theta \to \infty$, the NB reduces to Poisson. A small estimated θ indicates severe overdispersion; a large θ indicates the extra dispersion is modest.

a) Fit the negative binomial model with the same formula:

```
m_nb = glm.nb(art ~ fem + mar + kid5 + phd + ment,
              data = bioChemists)
summary(m_nb)
```

In a comment, compare the coefficient on `ment` to the Poisson estimate from the Poisson model above. Has it changed substantially? Report the estimated overdispersion parameter `theta` from the NB output. Is the overdispersion modest or severe?

b) Compare model fit using AIC:

```
AIC(m_pois, m_nb)
```

In a comment, report both AIC values. Which model has the lower AIC? Recall from earlier in the course that AIC penalizes model complexity, so a lower AIC for the NB model (which has one additional parameter) means the improvement in fit outweighs the added complexity. What does this comparison imply: is overdispersion a problem worth addressing for this dataset?

c) Compute predicted article counts for male vs. female researchers, holding all other variables at their sample means:

```
predictions(m_nb, newdata = datagrid(fem = c("Men", "Women")))
```

In a comment, report the predicted number of articles for men and women (with confidence intervals). How large is the gender gap in predicted publications? Is this difference statistically distinguishable given the uncertainty intervals?

d) Write a short summary paragraph as a comment in your R script (4–6 sentences). Cover all of the following: (1) whether Poisson regression is adequate for this dataset or whether the negative binomial is needed, and why; (2) the interpretation of the `ment` incidence rate ratio — what does mentor productivity tell us about student productivity?; (3) which predictors are statistically significant in the negative binomial model; (4) one substantive conclusion about the factors driving publication productivity among PhD students in biochemistry.

2 Part 2: Take-Home (Survival Analysis)

We now analyze time-to-event data using the `lung` dataset from the `survival` package, which contains survival times for patients with advanced lung cancer from the North Central Cancer Treatment Group. The outcome of interest is time from enrollment to death. Some patients are still alive at the end of the study — they are **right-censored**. Recall from class that ignoring censoring (e.g., dropping censored observations or treating them as events) produces biased estimates, which motivates the use of survival models.

Load the required packages (`survival`, `broom`, `ggplot2`, `marginaleffects`) and the `lung` dataset from `survival`.

Key variables in `lung`:

- `time` — survival time in days
- `status` — censoring status (1 = censored, 2 = dead)
- `age` — age in years
- `sex` — 1 = male, 2 = female
- `ph.ecog` — ECOG performance score (0 = fully active, . . . , 4 = bedridden)
- `ph.karno` — Karnofsky performance score rated by physician (0 = dead, 100 = normal)
- `wt.loss` — weight loss in last six months (pounds)

Note: The `status` variable uses 1 = censored and 2 = dead, but the `Surv()` function expects 0 = censored and 1 = event. Recode it before fitting any models (e.g. `lung$dead = lung$status - 1`).

2.1 Kaplan-Meier survival curves

a) Explore the data. In a comment, report the total number of observations, the number of events (deaths), and the number of censored cases. What proportion of patients are censored? Is this a lot or a little? Think about what this means: the censored patients' true survival times are unknown but at least as long as their observed times.

b) Estimate the overall Kaplan-Meier survival curve using `survfit()` with formula `Surv(time, dead) ~ 1`. The `summary()` output shows, at each event time, the number at risk, the number of events, the estimated survival probability, and the confidence interval. In a comment, report the estimated median survival time. What does this number mean in plain language?

c) Estimate separate Kaplan-Meier curves by sex using `survfit()` and plot them with `ggplot2`. Hint: use `broom::tidy()` to convert the `survfit` object to a data frame, then plot with `geom_step()` and `geom_ribbon()` for confidence intervals. Save the plot as a PDF. Also run a log-rank test using `survdiff()`. In a comment, describe what you see: which group survives longer? Does the confidence interval for the two groups overlap? Report the log-rank test p-value and explain what it tests (whether the survival curves are statistically different).

2.2 Cox proportional hazards model

a) Fit a Cox proportional hazards model predicting survival from `age`, `sex`, and `ph.ecog` using `coxph()` from `survival`. The output shows both raw coefficients (log-hazard scale) and exponentiated coefficients (hazard ratios). In a comment, report and interpret the hazard ratio for `sex`. Recall from class: a hazard ratio below 1 means lower hazard (longer survival), above 1 means higher hazard (shorter survival). What does the hazard ratio for sex tell us about survival differences between men and women? Is it statistically significant?

b) Interpret the hazard ratio for `ph.ecog`. In a comment, explain what a one-unit increase in ECOG performance score (i.e., moving toward worse physical functioning) does to the hazard of death. Express this as a percentage change (e.g., "X% higher/lower hazard").

c) The Cox model assumes **proportional hazards**: the effect of each covariate is constant over time. Test this assumption using `cox.zph()`. In a comment, report the p-value for each covariate and the global test. A significant p-value suggests the proportional hazards assumption is violated for that variable. Do any variables violate the assumption? If so, what would this mean in substantive terms (e.g., the effect of age changes over the course of the disease)?

d) Write a short summary paragraph as a comment in your R script (4–6 sentences). Cover: (1) whether the Kaplan-Meier analysis suggested survival differences by sex; (2) which predictors are significant in the Cox model and the direction of their effects; (3) whether the proportional hazards assumption holds; (4) one substantive conclusion about factors predicting lung cancer survival.

3 Data Sources

- British Electoral Panel Study: `BEPS` dataset in the `carData` R package (`library(carData); data(BEPS)`)
- PhD biochemist publications: `bioChemists` dataset in the `pscl` R package (`library(pscl); data(bioChemists)`)
- Lung cancer survival: `lung` dataset in the `survival` R package (`library(survival); data(lung)`)

4 Submission

Commit your file to your GitHub repository before the deadline. Put it in a separate folder, e.g. `assignment9`. Make sure your repository is public so I can access it.

Your R script should:

- Be well-organized with clear section headers (using comments)
- Include all code needed to reproduce your analysis
- Include your answers and interpretations as comments
- Save any plots to files (using `pdf()`/`dev.off()` or `ggsave()`)
- Run without errors from top to bottom