Week 5 Exercises

Note: The important **prospectus** is due on Oct. 1, so this homework is more brief than other weeks.

Exercise 1 {marginaleffects}

Read Arel-Bundock, Greifer, and Heiss (2024).[1] Write your own cheatsheet for the {marginaleffects} package. *At a minimum*, your cheatsheet should describe commonly used functions their commonly used arguments.

For inspiration: What are the important functions? What are the important arguments? What are good default practices—and how do these deviate from {marginaleffects} defaults? You can stick closely to the perspective in the notes or deviate far from that perspective.

No solution intended. Answers will vary.

Exercise 2 de Kadt and Grzymala-Busse (2025)

Read de Kadt and Grzymala-Busse (2025), especially Section 3.1, pp. 16-18. There (p. 17), they write:

> If the researcher firmly believes they are not engaged in counterfactual reasoning—they purely want to 'describe the data'—then the use of multivariate regression is itself a peculiar choice.

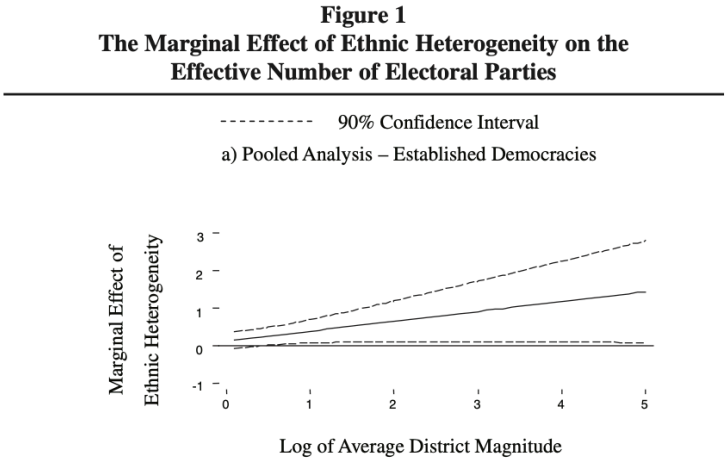
What do you make of their argument about muliple regression and control variables on pp. 16-18? Are control variables *always* about causal inference? Are control variables ever useful for description? Explain and use examples.

No solution intended. Answers will vary.

[1] There is also a book at marginaleffects.com/chapters/who.html with lots of examples and case studies.

Exercise 3 Clark and Golder (2006)

Use {marginaleffects} to reproduce *the spirit of*[2] Figure 1 on p. 701 of Clark and Golder (2006), which shows the effect of the effective number of ethnic groups on the effective number of ethnic parties as district magnitude varies.



The R code below gets you started.

```
# load data
cg <- crdata::cg2006

# fit model; "Established Democracies 1946-2000" model in Table 2 on p. 698
f <- enep ~ eneg*log(average_magnitude) + eneg*upper_tier + en_pres*proximity
fit <- lm(f, data = cg)
```

Solution

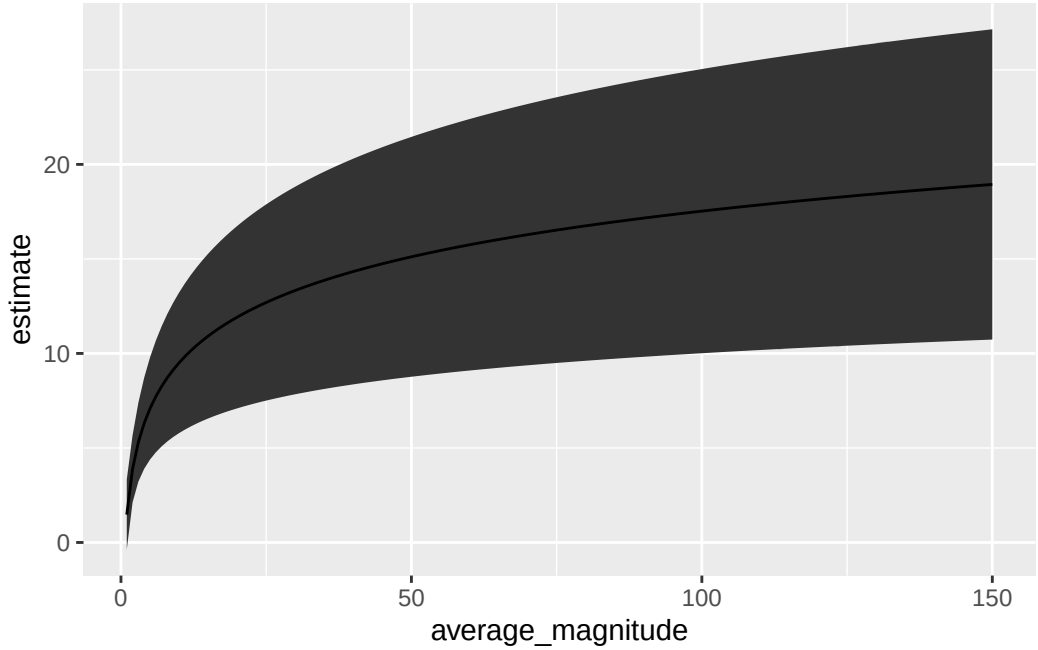
The code below serves as a minimal example–I tried to keep it as simple as possible. There are many things you might choose to do differently or better.

[2] By "the spirit of," I mean that you should feel free to change the mostly arbitrary features of the plot, like (1) the comparison (the authors use the instantaneous marginal effect), (2) the scale of the x-axis (the authors use the natural log), (3) the values of the other covariates, (4) average case or observed values, etc.

```r
# load packages
library(marginaleffects)

# compute first difference for a range of magnitudes
fd <- comparisons(fit,
                  variables = list(eneg = "minmax"),
                  newdata = datagrid(average_magnitude = 1:150,
                                     upper_tier = 0))

# plot
ggplot(fd, aes(x = average_magnitude, y = estimate,
               ymin = conf.low, ymax = conf.high)) +
  geom_ribbon() +
  geom_line()
```



The plot shows how the expected number of electoral parties (ENEP) changes when moving from the minimum to the maximum level of the effective number of ethnic groups (ENEG) as district magnitude varies from 1 to 150, fixing `upper_tier = 0` and holding other covariates at their means. For each value of district magnitude, the coefficient estimates are used to calculate this first difference and 95% confidence interval

Details

```r
fd <- comparisons(fit,
                  variables = list(eneg = "minmax"),
                  newdata = datagrid(average_magnitude = 1:150,
                                     upper_tier = 0))
```

- `comparisons(fit, ...)` calculates contrasts (differences in predictions) for a fitted model.
- `variables = list(eneg = "minmax")` specifies a **discrete contrast**: predict the outcome when `eneg` is set to its observed minimum, then to its observed maximum, and take the difference.
- `newdata = datagrid(...)` defines the evaluation dataset:
  - `average_magnitude = 1:150` generates 150 rows, one for each magnitude value between 1 and 150.
  - `upper_tier = 0` fixes to "no upper tier."
  - All other covariates are set to their typical values (means in this case).
- For each row in this grid, the model is used to predict the outcome under `eneg = min` and `eneg = max`, and their difference.

```r
ggplot(fd, aes(x = average_magnitude, y = estimate,
               ymin = conf.low, ymax = conf.high)) +
  geom_ribbon() +
  geom_line()
```

The plot answers: *"At each value of district magnitude between 1 and 150, in systems without an upper tier and other covariates at their mean, how much larger is the expected number of ethnic parties in the most ethnically diverse setting compared to the least?"* The solid line shows the estimated effect, and the ribbon shows the 90% around that estimate.

Note: Because this is a linear model, the "other covariates" only matter if they are included in interactions with variables of interest.

Exercise 4 `bioChemists`

Use one of our count models (Poisson, NB, and variants) to defend a useful/interesting/informative *descriptive* claim the number of articles that biochemistry graduate students produced during

the last three years. The `bioChemists` dataset in the {pscl} package has the relevant outcome `art`, as well as a number of potential covariates. See `?pscl::bioChemists` for the details.

1. With the "peculiar choice" argument of de Kadt and Grzymala-Busse (2025) in mind, select an appropriate set of control variables.
2. Use AIC and/or BIC to select a good model from our menu of options (Poisson/NB; ZI w/ covariates, constant ZI, no ZI; polynomials; interactions).
3. Use {marginaleffects} to compute quantities of interest for all, some, or the best of these models.

```
# load data
biochem <- pscl::bioChemists
```

No solution intended. Answers will vary.

Exercise 5 Hultman, Kathman, and Shannon (2013)

Hultman, Kathman, and Shannon (2013) use a negative binomial regression model.Skim through their paper to understand the application and the variables. You can find out more about their data and useful details with `?crdata::hks2013`.

1. Does a zero-inflated negative binomial model better fit their data? Consider a model with constant zero inflation and zero-inflation that depends on covariates Z. The Z variables can be the same as the X, or different.
2. Compute a substantively meaningful effect of UN troops on the number of civilians killed. How does the estimate for this quantity of interest change between their negative binomial regression compare to the alternative models you consider?

> ⚠ Warning
>
> The `MASS::glm.nb()` model has a hard time converging for this data. See the `control` options that I use below for the negative binomial model.

```
# load data
hks <- crdata::hks2013

# estimate models
f <- osvAll ~ troopLag + policeLag + militaryobserversLag +
  brv_AllLag + osvAllLagDum + incomp + epduration +
  lntpop
```

```
# replicates model 1 in table 1 on p. 884 of HKS
fit <- MASS::glm.nb(f, data = hks,
                    init.theta = 5,
                    control = glm.control(epsilon = 1e-12,
                                          maxit = 2500,
                                          trace = FALSE))
```

```
# load packages
library(glmmTMB)

# zinb w/o covariates
fit_zi0 <- glmmTMB(f, ziformula = ~ 1, data = hks, family = nbinom2)

# zinb w/ same covariates as nb portion
fit_zi1 <- glmmTMB(f, ziformula = update(f, NULL ~ .), data = hks, family = nbinom2)

# compare
BIC(fit, fit_zi0, fit_zi1)
```

```
        df      BIC
fit     10 12602.59
fit_zi0 11 12610.82
fit_zi1 19 11581.57
```

The BIC *strongly* infers the zero-inflated model with covariates modeling the zero inflation.

We can use use `avg_comparisons()` to compute the average change in expected civilian casualties as the number of UN troops (in 1000s) moves from 0 to it's mean (≈ 0.7).

```
bind_rows(
  "NB" = avg_comparisons(fit, variables = list(troopLag = c(0, .7))),
  "constant ZINB" = avg_comparisons(fit_zi0, variables = list(troopLag = c(0, .7))),
  "modeled ZINB" = avg_comparisons(fit_zi1, variables = list(troopLag = c(0, .7))),
  .id = "model") |>
  select(model, estimate, std.error) |>
  tinytable::tt()
```

For the negative binomial and zero-inflated negative binomial models with constant zero-inflation, we obtain absurdly large estimates and SEs. For the zero-inflated model with covariates, we get a reasonable estiamte that increasing the troops from zero to their average level in the data decreases civilian casaulties by about 1,000, give or take 1,300 or so.

model	estimate	std.error
NB	-2159660.791	3028340.011
constant ZINB	-2159703.782	3172716.794
modeled ZINB	-1065.135	1302.757

These estimates make it clear that *something weird* is happening with these estimates. It turns out that their are some *extremely* large counts in these data.

```
summary(hks$osvAll)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
     0.00      0.00      0.00     69.54      0.00 145844.00
```

These extremely large counts push the overdispersion parameter *very close to zero*. When the overdispersion parameter is very close to zero, the mean is very large and the SE is very large.

```
summary(fit)
```

```
Call:
MASS::glm.nb(formula = f, data = hks, control = glm.control(epsilon = 1e-12,
    maxit = 2500, trace = FALSE), init.theta = 0.05919426018,
    link = log)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -9.2367536  0.8750837 -10.555   <2e-16 ***
troopLag             -0.5304466  0.0623706  -8.505   <2e-16 ***
policeLag            -9.9025813  1.0421355  -9.502   <2e-16 ***
militaryobserversLag 21.7618689  1.3017670  16.717   <2e-16 ***
brv_AllLag            0.0007062  0.0005804   1.217    0.224
osvAllLagDum          2.1773614  0.1762377  12.355   <2e-16 ***
incomp                2.3793901  0.1871308  12.715   <2e-16 ***
epduration           -0.0005591  0.0013581  -0.412    0.681
lntpop                0.7031066  0.0726654   9.676   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for Negative Binomial(0.0592) family taken to be 1)

    Null deviance: 2862.1  on 3745  degrees of freedom
Residual deviance: 1834.7  on 3737  degrees of freedom
AIC: 12540

Number of Fisher Scoring iterations: 1

              Theta:  0.05919
          Std. Err.:  0.00228

 2 x log-likelihood:  -12520.30800
```

Modeling the zero-inflation allows the model to adequately capture the overdispersion while keeping θ to a reasonable value (i.e., about 0.19 rather than 0.06).

```
summary(fit_zi1)
```

```
 Family: nbinom2  ( log )
Formula:
osvAll ~ troopLag + policeLag + militaryobserversLag + brv_AllLag +
    osvAllLagDum + incomp + epduration + lntpop
Zero inflation:
~troopLag + policeLag + militaryobserversLag + brv_AllLag + osvAllLagDum +
    incomp + epduration + lntpop
Data: hks

      AIC       BIC    logLik -2*log(L)  df.resid
  11463.2   11581.6   -5712.6   11425.2      3727


Dispersion parameter for nbinom2 family (): 0.192

Conditional model:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -4.2935452  0.9193405  -4.670 3.01e-06 ***
troopLag             -0.2973496  0.0953143  -3.120  0.00181 **
policeLag            -7.5250236  1.2462274  -6.038 1.56e-09 ***
militaryobserversLag 13.0228012  1.1210829  11.616  < 2e-16 ***
brv_AllLag            0.0003716  0.0002883   1.289  0.19745
osvAllLagDum          0.0637379  0.1839917   0.346  0.72903
incomp                1.9312635  0.2052219   9.411  < 2e-16 ***
```

```
epduration           -0.0096236  0.0012134  -7.931 2.17e-15 ***
lntpop                0.5675421  0.0754175   7.525 5.26e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Zero-inflation model:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           10.915372   1.232630   8.855  < 2e-16 ***
troopLag               0.146006   0.080162   1.821   0.0685 .
policeLag              3.738534   2.064429   1.811   0.0702 .
militaryobserversLag  -4.416012   1.909655  -2.312   0.0208 *
brv_AllLag            -0.023187   0.012906  -1.797   0.0724 .
osvAllLagDum         -20.318188 821.539894  -0.025   0.9803
incomp                -1.647894   0.251947  -6.541 6.13e-11 ***
epduration            -0.008084   0.001533  -5.275 1.33e-07 ***
lntpop                -0.596124   0.096498  -6.178 6.51e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise 6 Radean and Beger (2025)

Read Radean and Beger (2025). The authors make an important point. They argue that reporting only the average effect can be misleading; it can obscure the wide variation in effects across observed values. Instead, they advocate for a case-centered approach, where researchers compute and report effects for each observation.

Use the data and model from Russett and Oneal (2001) in the {crdata} package to illustrate the relative value of a single summary effect (e.g., the average difference across all observations) and the approach recommended by Radean and Beger (2025).

```r
# load data
ro <- crdata::ro2001

# glm version of their gee on pp. 314
f <- dispute ~ allies + lcaprat2 + contiguity + dem.lo + logdstab + power
fit <- glm(f, family = "binomial", data = ro)

# example quantity of interest
avg_comparisons(fit,  variables = list(dem.lo = c(-10, 10)))
```

```
 Estimate Std. Error     z Pr(>|z|)     S   2.5 %  97.5 %
  -0.0605    0.00272 -22.3   <0.001 362.3 -0.0658 -0.0552

Term: dem.lo
Type: response
Comparison: 10 - -10
```

No solution intended. Answers will vary.

The idea is to use `avg_comparisons()` to compute a single summary effect and use `comparisons()` to show the richness that gets lost when applying the `avg_*()`.

Exercise 7 Berry, DeMeritt, and Esarey (2009)

Berry, DeMeritt, and Esarey (2009) have a large dataset and consider a number of polynomials and interactions in their probit model. Using their Model 1 as a starting point, use the AIC and/or BIC to find the best model along two dimensions.

1. **Link Functions**: The authors use `probit`, but other options from within `glm()` are `logit`, `cauchit`, `log`, and `cloglog`.[3]. `family = binomial()` uses logit by default; `family = binomial(link = "probit")` uses probit instead. Fit their Model 1 using each of these link functions and use the BIC and/or AIC to find the best. Explain the results.
2. **Interactions and Polynomials**: The authors consider two models. Both have polynomials; one model has interactions as well, the other does not. Consider a simpler model without interactions *or* polynomials. And consider higher-order interactions, deeper interactions, or both. Make sure to use R formula syntax for compact and clear representation. Can you justify a more complicated model that their Model 1 with the IC? Explain the results.

You can find the `scobit.dta` data set here.

[3]From `?family`: "the binomial family the links `logit`, `probit`, `cauchit`, (corresponding to logistic, normal and Cauchy CDFs respectively) and `log` and `cloglog` (complementary log-log)."

TABLE 1 Probit Models of 1984 Presidential Voting Turnout

	Dependent Variable: Pr(*vote*)	
	(1)	(2)
Independent Variable	**Model with Product Terms (i.e., Nagler Specification)**	**Model without Product Terms (i.e., Wolfinger & Rosenstone Specification)**
closing date	0.0006	−0.0078*
	(0.0037)	(0.0004)
education	0.2645*	0.1819*
	(0.0416)	(0.0144)
$education^2$	0.0051	0.0123*
	(0.0042)	(0.0014)
age	0.0697*	0.0697*
	(0.0013)	(0.0013)
age^2	−0.0005*	−0.0005*
	(0.0000)	(0.0000)
south	−0.1155*	−0.1159*
	(0.0110)	(0.0110)
gubernatorial election	0.0034	0.0034
	(0.0116)	(0.0116)
closing date ∗ *education*	−0.0032*	
	(0.0015)	
closing date ∗ $education^2$	0.00028	
	(0.00015)	
constant	−2.7431*	−2.5230*
	(0.1074)	(0.0486)
N	99,676	99,676
Log-likelihood	−55815.28	−55818.03

Standard errors in parentheses.
* $p \leq .05$.

```
# code to load and clean the data
scobit <- haven::read_dta("data/scobit.dta") %>%
  filter(newvote != -1)

# fit author's model 1
f <- newvote ~ poly(neweduc, 2)*closing + poly(age, 2) + south + gov
fit1 <- glm(f, family = binomial, data = scobit)

# fit wildly complicated model (51 parameters)
f <- newvote ~ (poly(neweduc, 2) + closing + poly(age, 2) + south + gov)^3
fit2 <- glm(f, family = binomial, data = scobit)

# evaluate w/ bic
BIC(fit1, fit2)
```

```
     df       BIC
fit1 10 111664.2
fit2 51 111707.8
```

Part 1

The solution doesn't need to use this approach, but this is a good time to use functions, list, and `map()`.

```
# a function to fit their model 1 for a supplied link function
fit_model <- function(link) {
  f <- newvote ~ poly(neweduc, 2)*closing + poly(age, 2) + south + gov
  glm(f, family = binomial(link = link), data = scobit)
}

# links to consider
links <- c("logit", "probit", "cauchit", "cloglog")

# fit each model; compute BIC; print neatly
set_names(links) |>
  map(fit_model) |>
  map(BIC) |>
  as_tibble()
```

```
# A tibble: 1 x 4
    logit  probit cauchit cloglog
    <dbl>   <dbl>   <dbl>   <dbl>
1 111664. 111746. 111665. 112089.
```

The code above uses a compact functional approach in R to compare alternative link functions for Berry et al.'s Model 1. A helper function fit the model for each candidate link (`logit`, `probit`, `cauchit`, and `cloglog`). Then `purrr::map()` fits all models, computes the BIC, and prints the BIC in a tidy table. Comparing BIC values shows that the logit and cauchit links perform almost identically. Both perform slightly better than probit, while the cloglog link is clearly worse. Although the authors' use probit, the data provide some support for alternative link functions, particularly logit or cauchit.

Part 2

Answers will vary.

References

Arel-Bundock, Vincent, Noah Greifer, and Andrew Heiss. 2024. "How to Interpret Statistical Models Using Marginaleffects for R and Python." *Journal of Statistical Software* 111 (9). https://doi.org/10.18637/jss.v111.i09.

Berry, William D., Jacqueline H. R. DeMeritt, and Justin Esarey. 2009. "Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?" *American Journal of Political Science* 54 (1): 248–66. https://doi.org/10.1111/j.1540-5907.2009.00429.x.

Clark, William Roberts, and Matt Golder. 2006. "Rehabilitating Duverger's Theory." *Comparative Political Studies* 39 (6): 679–708. https://doi.org/10.1177/0010414005278420.

de Kadt, Daniel, and Anna Grzymala-Busse. 2025. "Good Description." SocArXiv. https://doi.org/10.31235/osf.io/e74a9_v1.

Hultman, Lisa, Jacob Kathman, and Megan Shannon. 2013. "United Nations Peacekeeping and Civilian Protection in Civil War." *American Journal of Political Science* 57 (4): 875–91. https://doi.org/10.1111/ajps.12036.

Radean, Marius, and Andreas Beger. 2025. "Not-so-Average After All: Individual Vs. Aggregate Effects in Substantive Research." *Journal of Peace Research*. https://repository.essex.ac.uk/40373/.

Russett, Bruce, and John R. Oneal. 2001. *Triangulating Peace: Democracy, Interdependence, and International Organizations*. New York: W. W. Norton & Company.