An Application of Ordinal Regression to Attitudes about the Risks of Vaccines

Lisa Wilson

Fall 2019

Background and Data

In recent years in the United States, anti-vaccine sentiment, including the unsubstantiated belief that childhood vaccines can cause autism, has contributed to an increased number of unvaccinated children and to outbreaks of previously well-controlled diseases such as measles.[1,2] In a survey conducted in May and June 2016, the Pew Research Center asked respondents a number of questions related to their attitudes toward vaccines.[3] One of these questions asked, "Thinking about childhood vaccines for measles, mumps and rubella (MMR) how would you rate the risk of side effects?" The possible responses were very high, high, medium, low, and very low. Because the possible answers are ordered, this question makes an appropriate response variable for an ordinal regression, where other survey questions could be explored as possible covariates to see which factors have a significant effect or influential predictive ability on belief about vaccine risk. The total sample size for the survey was 4563 respondents, and after data cleaning to remove incomplete entries for the questions of interest, 391 observations remained to be analyzed.

Fitting a full model

Out of over 50 possible covariates, 15 were identified for fitting an initial "full model":

- Whether the respondent had been vaccinated as a child
- Rating questions (all with ordered categorical responses):
  - How well do they believe medical scientists understand risks and benefits of vaccines
  - How many scientists do they believe say vaccines are safe
  - How closely do they follow news about childhood vaccines
- Demographic questions (all with categorical responses): region of the U.S., age, sex, education, race, whether the respondent is a born-again or evangelical Christian, party affiliation, political ideology (i.e., conservative, moderate, or liberal), income level, level of internet use, whether the respondent uses social media

A logistic-link ordinal regression model was fit with these 15 covariates and the response variable about vaccine risks, where "very low" was the reference category.

	Value	Std. Error	t value	p value
Vaccinated	-1.1037	0.4802	-2.2981	0.0216
Understand: Not too well	0.7796	0.3847	2.0265	0.0427
Understand: Very well	-1.2903	0.2380	-5.4214	0.0000
Scientists: Almost all	-1.0995	0.3817	-2.8802	0.0040
Northeast	0.9031	0.3337	2.7064	0.0068
Asian	-2.4760	1.2204	-2.0288	0.0425
Black	1.0126	0.4491	2.2547	0.0242
Evangelical	0.5180	0.2226	2.3270	0.0200
Conservative	0.5800	0.2389	2.4275	0.0152

The above table shows the logistic-link ordinal regression output for factors that were significant (p-value < 0.05). A negative coefficient indicates someone in that category is less likely than the reference group to believe vaccine side effects are greater than very low, or, in other words, that someone in that category is less likely to believe that some non-neglible vaccine risk exists. According to this model, the significant factors with negative coefficients are the following: whether the respondent had been vaccinated, if the respondent believes medical scientists understand the risks and benefits of vaccines very well, if the respondent believes almost all scientists say vaccines are safe, and if the respondent is Asian. A positive coefficient indicates someone in that category is more likely than the reference group to believe vaccine side effects are greater than very low. The significant factors with negative coefficients are the following: if the respondent believes medical scientists understand the risks and benefits of vaccines not too well, is from the Northeast, is Black, is evangelical, or is politically conservative.

Interpreting odds ratios

Because the full model was fit with a logistic link, exponentiating the coefficients from the previous table results in odds ratios that can be more easily interpreted. For example, holding other covariates constant,

- For someone answering that medical scientists understand the risks & benefits of vaccines *very well*, the odds of believing some non-neglible vaccine risk exists (i.e., vaccine risks are are greater than very low) is 0.275 times that of someone answering *fairly well*.
- For a person from the *Northeast*, the odds of believing some vaccine risk exists is 2.467 times that of a person from the *Midwest*.
- For a *born-again or evangelical* Christian, the odds of believing some vaccine risk exists is 1.679 times that of someone who is not.

Finding a predictive model

Model selection tools were next used to find a model that could be used to predict how respondents would rate the risk of vaccines (although inference on specific covariates is no longer valid after model selection). Incorporating forward and backward selection, the `R` function `stepAIC` identified a model with the following covariates as having the lowest AIC:

- Survey questions: whether the respondent was vaccinated, how well do they believe medical scientists understand risks and benefits of vaccines, how many scientists do they believe say vaccines are safe, how closely do they follow news about childhood vaccines
- Demographics: region, race, whether evangelical, political ideology, level of internet use

The regression output for this model is shown below:

	Value	Std. Error	t value	p value
Don't remember	-0.2057	0.6479	-0.3175	0.7508
Vaccinated	-1.3295	0.4677	-2.8426	0.0045
Understand: Not at all well	0.8980	0.6200	1.4482	0.1475
Understand: Not too well	0.7562	0.3753	2.0150	0.0439
Understand: Very well	-1.2931	0.2349	-5.5038	0.0000
Scientists: Almost all	-1.1613	0.3695	-3.1432	0.0017
Scientists: Almost none	-1.8041	1.1298	-1.5969	0.1103
Scientists: Fewer than half	0.9501	0.6071	1.5650	0.1176
Scientists: More than half	-0.5764	0.3671	-1.5703	0.1163
News: Not at all closely	-0.4561	0.3671	-1.2427	0.2140
News: Not too closely	-0.4223	0.2312	-1.8267	0.0677
News: Very closely	0.4822	0.3197	1.5083	0.1315
Northeast	0.9718	0.3244	2.9961	0.0027
South	-0.0885	0.2523	-0.3508	0.7258
West	0.4540	0.3037	1.4951	0.1349

Asian	-2.4157	1.2413	-1.9461	0.0516
Black	0.9372	0.3983	2.3532	0.0186
Mixed race	0.5275	0.4360	1.2099	0.2263
Other race	-0.6288	0.4921	-1.2780	0.2013
Evangelical	0.4720	0.2163	2.1820	0.0291
Conservative	0.4437	0.2206	2.0115	0.0443
Liberal	0.0418	0.3192	0.1310	0.8958
Internet: Medium	-0.4753	0.3105	-1.5308	0.1258
Internet: High	0.1614	0.2637	0.6121	0.5404

Trying different link functions, the logistic link model has the lowest AIC at 969.433, compared to the probit link (AIC of 974.697) or the log-log link (AIC of 986.557).

As an example, this model was used to predict the responses for the "most typical" respondent. The characteristics of the most typical respondent were identified by selecting the mode response for each covariate (although no observation exists in the data where someone has all of these characteristics).

Accordingly, the "most typical" respondent in this sample:

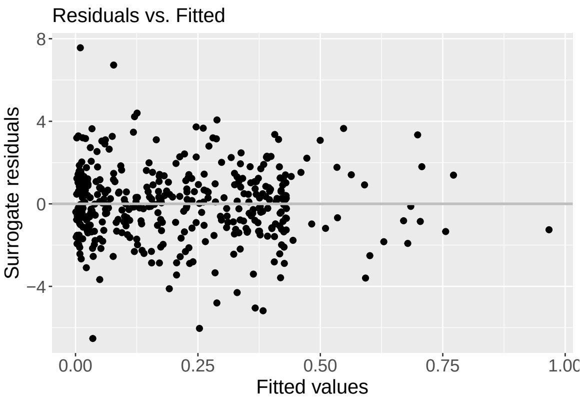
- had childhood vaccines
- is a 50 to 64 year old white female from the South with at least a college degree who makes at least $75,000
- is moderate, Republican, and not evangelical
- uses the internet frequently and uses social media
- believes medical scientists understand vaccine risks and benefits very well
- believes almost all medical scientists think vaccines are safe
- follows vaccine news somewhat closely

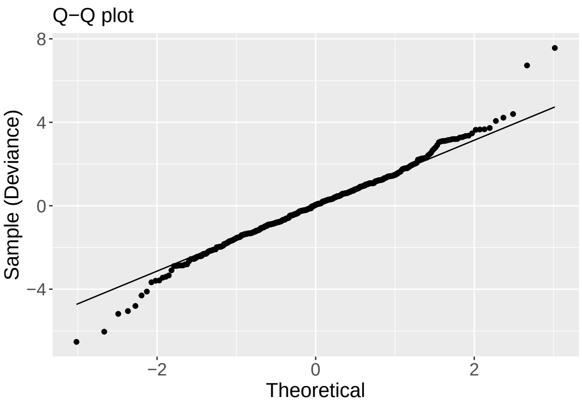
The model then predicts this person would rate the risk of side effects from childhood vaccines with the following probabilities:

Very low	Low	Medium	High	Very high
0.6893	0.2441	0.05489	0.008271	0.003445

The most typical respondent in this sample is, according to the predictive model, most likely to identify the risk of side effects from vaccines as very low, with a predicted probability of 0.6893. There is a predicted probability of only 0.0034 that this respondent would rate vaccine risks as very high.

Residual analysis





Using the package `sure` to obtain surrogate residuals, the residual vs. fitted values plot and the Q-Q plot seem to indicate that the predictive model fits the data well. There does not appear to be a clear pattern in the residuals vs. fitted plot, and the residuals appear to be symmetric around zero. Likewise, the Q-Q plot does not show evidence of non-normality in the errors, just some indication of longer left and right tails than would be expected.

Collapsing response categories

Finally, the response categories were collapsed to see if doing so improves model performance. In fact, if the response categories are collapsed to High, Medium, and Low and the same predictive model is fit, the AIC reduces to 576.358 from 969.433.

The predicted probabilities of each response category for the most typical respondent then become

Low	Medium	High
0.9222	0.0644	0.0134

Conclusion

Based on this survey data and an ordinal regression analysis, whether a person was vaccinated and how much they trust medical scientists seem to predict how they rate the risks of vaccines. Influential demographic data for this prediction seems to be region, race, political ideology, evangelicalism, and internet use. Furthermore, as long as more specific response categories are not of interest (i.e., if the distinction between "very high" and "high" is not important), the predictive model can be improved by collapsing the response categories.

References

1. Leventhal, Jamie. "A quiet rise in unvaccinated children could put the U.S. at risk of outbreaks." PBS Newshour. October 12, 2018. https://www.pbs.org/newshour/health/a-quiet-rise-in-unvaccinated-children-could-put-the-u-s-at-risk-of-outbreaks.
2. "In Wake of Measles Outbreaks, CDC Updates 2019 Case Totals." American Academy of Family Physicians. October 9, 2019. https://www.aafp.org/news/health-of-the-public/20191009measlesupdt.html.
3. "American Trends Panel Wave 17." Pew Research Center. Accessed November 17, 2019. https://www.pewresearch.org/science/dataset/american-trends-panel-wave-17/. ["Download Dataset" near bottom of page.]

Code appendix

```r
## data import
# initial selection of relevant variables in usable form
library(foreign)

setwd("ST 623/")
prc <- data.frame(read.spss("W17_May16/ATP W17.sav"))
str(prc)
head(prc)

colnames(prc)

# BIO33A_W17 through BIO46_W17 are on vaccines
# LOCALD_W17 through LOCALF_W17, ENV1_W17 through ENV34_W17 are environmental
# Relevant demo: F_CREGION_FINAL, F_AGECAT_FINAL, F_SEX_FINAL, F_EDUCCAT_FINAL (+ another),
# F_HISP_RECRUITMENT, F_RACECMB_RECRUITMENT, F_RACETHN_RECRUITMENT, F_RELIG_FINAL, F_BORN_FINAL,
# F_RELCOM3CAT_FINAL, F_PARTY_FINAL, F_PARTYSUM_FINAL, F_INCOME_RECODE_FINAL
# F_IDEO_FINAL, F_SNSUSER_FINAL
prc_vacc <- prc[,c(16, 17, 25:27, 58, 60:62, 68:93, 286:314)]
head(prc_vacc)
summary(prc_vacc[,1])

# attached in email
write.csv(prc_vacc, "prc_vacc", row.names=FALSE)

library(MASS)
library(tidyverse)
library(sure)
library(pander)
library(gdata)
library(kableExtra)

## data cleaning
vacc <- read_csv("prc_vacc")
# get rid of confidence questions and how much have used alternative med
# so that na.omit doesn't nullify dataset
vacc_full <- data.frame(na.omit(vacc[,-c(1:2, 9)]))

for(i in 1:ncol(vacc_full)){
  vacc_full[,i] <- as.factor(vacc_full[,i])
}

vacc_full$BIO21_W17 <- relevel(vacc_full$BIO21_W17, "No, was not")
# vacc_full$F_PARTYSUM_FINAL <- relevel(vacc_full$F_PARTYSUM_FINAL, "Independent/No Lean")
vacc_full$F_IDEO_FINAL <- fct_collapse(vacc_full$F_IDEO_FINAL,
                                       Conservative = c("Very conservative", "Conservative"),
                                       Moderate = "Moderate",
                                       Liberal = c("Very liberal", "Liberal"))
vacc_full$F_IDEO_FINAL <- factor(vacc_full$F_IDEO_FINAL, levels(vacc_full$F_IDEO_FINAL)[c(3,1,2)])
vacc_full$F_INT_FREQCOMB_FINAL <- fct_collapse(vacc_full$F_INT_FREQCOMB_FINAL,
                                   Low = c("Never use the Internet",
                                           "Use the Internet less than once a month",
                                           "Use the Internet once a month",
                                           "Use the Internet once a week",
                                           "Use the Internet at least once a week but not every day"),
                                   Medium = c("Use the Internet about once a day",
                                              "Use the Internet a few times a day"),
                                   High = c("Use the Internet many times a day",
                                            "Use the Internet constantly"))
vacc_full$F_RACECMB_RECRUITMENT <- relevel(vacc_full$F_RACECMB_RECRUITMENT, "White")
vacc_full$F_BORN_FINAL <- factor(vacc_full$F_BORN_FINAL, levels(vacc_full$F_BORN_FINAL)[c(2,3)])
vacc_full$F_EDUCCAT_FINAL <- factor(vacc_full$F_EDUCCAT_FINAL,
                                    levels(vacc_full$F_EDUCCAT_FINAL)[c(2,3,1)])
vacc_full$BIO33A_W17 <- factor(vacc_full$BIO33A_W17, levels(vacc_full$BIO33A_W17)[c(6,2,3,1,5,4)])
vacc_full$BIO33B_W17 <- factor(vacc_full$BIO33B_W17, levels(vacc_full$BIO33B_W17)[c(5,1,3,2,6,4)])
vacc_full$BIO36_W17 <- relevel(vacc_full$BIO36_W17, "Some")
vacc_full$BIO42_W17 <- relevel(vacc_full$BIO42_W17, "Somewhat closely")
vacc_full$BIO43_W17 <- relevel(vacc_full$BIO43_W17, "Somewhat good job")

use_A <- c("BIO21_W17", "BIO33A_W17", "BIO38_W17", "BIO40_W17", "BIO42_W17", "F_CREGION_FINAL",
           "F_AGECAT_FINAL", "F_SEX_FINAL", "F_EDUCCAT_FINAL", "F_RACECMB_RECRUITMENT", "F_BORN_FINAL",
           "F_PARTYSUM_FINAL", "F_INCOME_RECODE_FINAL", "F_IDEO_FINAL", "F_INT_FREQCOMB_FINAL",
           "F_SNSUSER_FINAL")

vacc_A <- na.omit(vacc_full[, use_A])
refuse <- vacc_A %>%
  filter_all(any_vars(str_detect(., pattern = "Refused")))
vacc_A <- drop.levels(anti_join(vacc_A, refuse), reorder = FALSE)

## full model
modA <- polr(BIO33A_W17~., data = vacc_A)
ctableA <- coef(summary(modA))
pA <- round(pnorm(abs(ctableA[, "t value"]), lower.tail = FALSE) * 2, 4)
ctableA <- cbind(ctableA, "p value" = pA)
ctableA_abb <- ctableA[ctableA[,4] <= 0.05, ]
ctableA_abb <- ctableA_abb[-c(10:11),]
rownames(ctableA_abb) <- c("Vaccinated", "Understand: Not too well", "Understand: Very well",
                           "Scientists: Almost all", "Northeast", "Asian", "Black", "Evangelical",
                           "Conservative")

kable(round(ctableA_abb, 4)) %>%
  kable_styling(font_size = 10)

oddsA <- exp(ctableA_abb[,1])

## predictive model
stepAIC(polr(BIO33A_W17~., data=vacc_A))
stepAIC(polr(BIO33A_W17~., data=vacc_A, method = "probit"))
stepAIC(polr(BIO33A_W17~., data=vacc_A, method = "loglog"))

# model identified from stepAIC
modA_step <- polr(BIO33A_W17 ~ BIO21_W17 + BIO38_W17 + BIO40_W17 +
    BIO42_W17 + F_CREGION_FINAL + F_RACECMB_RECRUITMENT + F_BORN_FINAL +
    F_IDEO_FINAL + F_INT_FREQCOMB_FINAL, data = vacc_A)
# round(pnorm(abs(coef(summary(modA_step))[, "t value"]), lower.tail = FALSE) * 2, 3)
```

```
modAstep_log <- polr(BIO33A_W17 ~ BIO21_W17 + BIO38_W17 + BIO40_W17 +
    BIO42_W17 + F_CREGION_FINAL + F_RACECMB_RECRUITMENT + F_BORN_FINAL +
    F_IDEO_FINAL + F_INT_FREQCOMB_FINAL, data = vacc_A, method = "logistic")
modAstep_log.aic <- extractAIC(modAstep_log)[[2]]

modAstep_probit <- polr(BIO33A_W17 ~ BIO21_W17 + BIO38_W17 + BIO40_W17 +
    BIO42_W17 + F_CREGION_FINAL + F_RACECMB_RECRUITMENT + F_BORN_FINAL +
    F_IDEO_FINAL + F_INT_FREQCOMB_FINAL, data = vacc_A, method = "probit")
modAstep_probit.aic <- extractAIC(modAstep_probit)[[2]]

modAstep_loglog <- polr(BIO33A_W17 ~ BIO21_W17 + BIO38_W17 + BIO40_W17 +
    BIO42_W17 + F_CREGION_FINAL + F_RACECMB_RECRUITMENT + F_BORN_FINAL +
    F_IDEO_FINAL + F_INT_FREQCOMB_FINAL, data = vacc_A, method = "loglog")
modAstep_loglog.aic <- extractAIC(modAstep_loglog)[[2]]

ctableA2 <- coef(summary(modA_step))
pA2 <- round(pnorm(abs(ctableA2[, "t value"]), lower.tail = FALSE) * 2, 4)
ctableA2 <- cbind(ctableA2, "p value" = pA2)[-c(25:28),]
rownames(ctableA2) <- c("Don't remember", "Vaccinated", "Understand: Not at all well",
                        "Understand: Not too well", "Understand: Very well", "Scientists: Almost all",
                        "Scientists: Almost none", "Scientists: Fewer than half",
                        "Scientists: More than half",
                        "News: Not at all closely", "News: Not too closely", "News: Very closely",
                        "Northeast", "South", "West", "Asian",
                        "Black", "Mixed race", "Other race",
                        "Evangelical", "Conservative", "Liberal",
                        "Internet: Medium", "Internet: High")

kable(round(ctableA2, 4), longtable = TRUE) %>%
  kable_styling(font_size = 10)

# Using "mode" to predict
modA_pred <- predict(modAstep_log,
        newdata = data.frame(BIO21_W17 = "Yes, was vaccinated for the major childhood diseases",
                             BIO38_W17 = "Very well", BIO40_W17 = "Almost all",
                             BIO42_W17 = "Somewhat closely", F_CREGION_FINAL = "South",
                             F_AGECAT_FINAL = "50-64", F_SEX_FINAL = "Female",
                             F_EDUCCAT_FINAL = "College graduate+", F_RACECMB_RECRUITMENT = "White",
                             F_BORN_FINAL = "No, not born-again or evangelical Christian",
                             F_PARTYSUM_FINAL = "Rep/Rep Lean",
                             F_INCOME_RECODE_FINAL = "$75,000+",
                             F_IDEO_FINAL = "Moderate", F_INT_FREQCOMB_FINAL = "High",
                             F_SNSUSER_FINAL = "Social Media Users"),
        type = "p")

pander(modA_pred)

# mode observation doesn't exist in dataset
mode_data <- vacc_A %>%
  filter(BIO21_W17 == "Yes, was vaccinated for the major childhood diseases",
         BIO38_W17 == "Very well", BIO40_W17 == "Almost all", BIO42_W17 == "Somewhat closely",
         F_CREGION_FINAL == "South", F_AGECAT_FINAL == "50-64",
         F_SEX_FINAL == "Female", F_EDUCCAT_FINAL == "College graduate+",
         F_RACECMB_RECRUITMENT == "White",
         F_BORN_FINAL == "No, not born-again or evangelical Christian",
         F_PARTYSUM_FINAL == "Rep/Rep Lean",
         F_INCOME_RECODE_FINAL == "$75,000+",
         F_IDEO_FINAL == "Moderate",
         F_INT_FREQCOMB_FINAL == "High",
         F_SNSUSER_FINAL == "Social Media Users")

## Residual checking
# Obtain surrogate residuals
set.seed(101) # for reproducibility
sresA <- resids(modAstep_log)
fitA <- unique(c(modAstep_log$fit[,1], modAstep_log$fit[,2], modAstep_log$fit[,3],
                 modAstep_log$fit[,4], modAstep_log$fit[,5]))
fitA.samp <- sample(fitA, length(sresA))

# Residual-vs-fitted plot
ggplot() +
    geom_point(aes(x=fitA.samp, y=sresA), size = 2) +
    geom_abline(intercept = 0, slope = 0, color = "grey", size = 1) +
    labs(x = "Fitted values", y = "Surrogate residuals", title = "Residuals vs. Fitted") +
    theme(axis.title = element_text(size = 16), axis.text = element_text(size = 14),
        plot.title = element_text(size = 16))

# normal enough
ggplot() +
  geom_qq_line(aes(sample = sresA)) +
  geom_qq(aes(sample = sresA)) +
  labs(title="Q-Q plot", x="Theoretical", y="Sample (Deviance)") +
  theme(axis.title = element_text(size = 16), axis.text = element_text(size = 14),
        plot.title = element_text(size = 16))

## collapse categories and look at difference in residuals
vacc_A3 <- vacc_A
vacc_A3$BIO33A_W17 <- fct_collapse(vacc_A3$BIO33A_W17, High = c("Very high", "High"),
                                   Medium = "Medium", Low = c("Low", "Very low"))

modA_collapse <- polr(BIO33A_W17 ~ BIO21_W17 + BIO38_W17 + BIO40_W17 +
    BIO42_W17 + F_CREGION_FINAL + F_RACECMB_RECRUITMENT + F_BORN_FINAL +
    F_IDEO_FINAL + F_INT_FREQCOMB_FINAL, data = vacc_A3)

modA_c.aic <- extractAIC(modA_collapse)[[2]]

# Using "mode" to predict with collapsed categories
modA_c_pred <- predict(modA_collapse,
        newdata = data.frame(BIO21_W17 = "Yes, was vaccinated for the major childhood diseases",
                             BIO38_W17 = "Very well", BIO40_W17 = "Almost all",
                             BIO42_W17 = "Somewhat closely", F_CREGION_FINAL = "South",
                             F_AGECAT_FINAL = "50-64", F_SEX_FINAL = "Female",
                             F_EDUCCAT_FINAL = "College graduate+", F_RACECMB_RECRUITMENT = "White",
                             F_BORN_FINAL = "No, not born-again or evangelical Christian",
                             F_PARTYSUM_FINAL = "Rep/Rep Lean",
                             F_INCOME_RECODE_FINAL = "$75,000+",
                              F_IDEO_FINAL = "Moderate", F_INT_FREQCOMB_FINAL = "High",
                              F_SNSUSER_FINAL = "Social Media Users"),
        type = "p")

pander(modA_c_pred)
```