HW 10

Sumit Gupta

November 15, 2017

1. Using the anes_2008tr.csv dataset in Course Resources, model vote_rep (whether the respondent voted Republican in the last election) as a function of age, race, income, and ideology.

```
data <- read.csv("anes_2008tr.csv", header = T, sep = ",")
data[1:2, 1:7]
```

```
##   age race_white gender_male education income ideology_con partyid_rep
## 1  35          0           1         5      4            4           5
## 2  58          1           0         3      3            6           1
```

```
dim(data)
```

```
## [1] 2322    9
```

```
invlogit <- function(x) exp(x) / (1+exp(x))
library(stargazer)
```

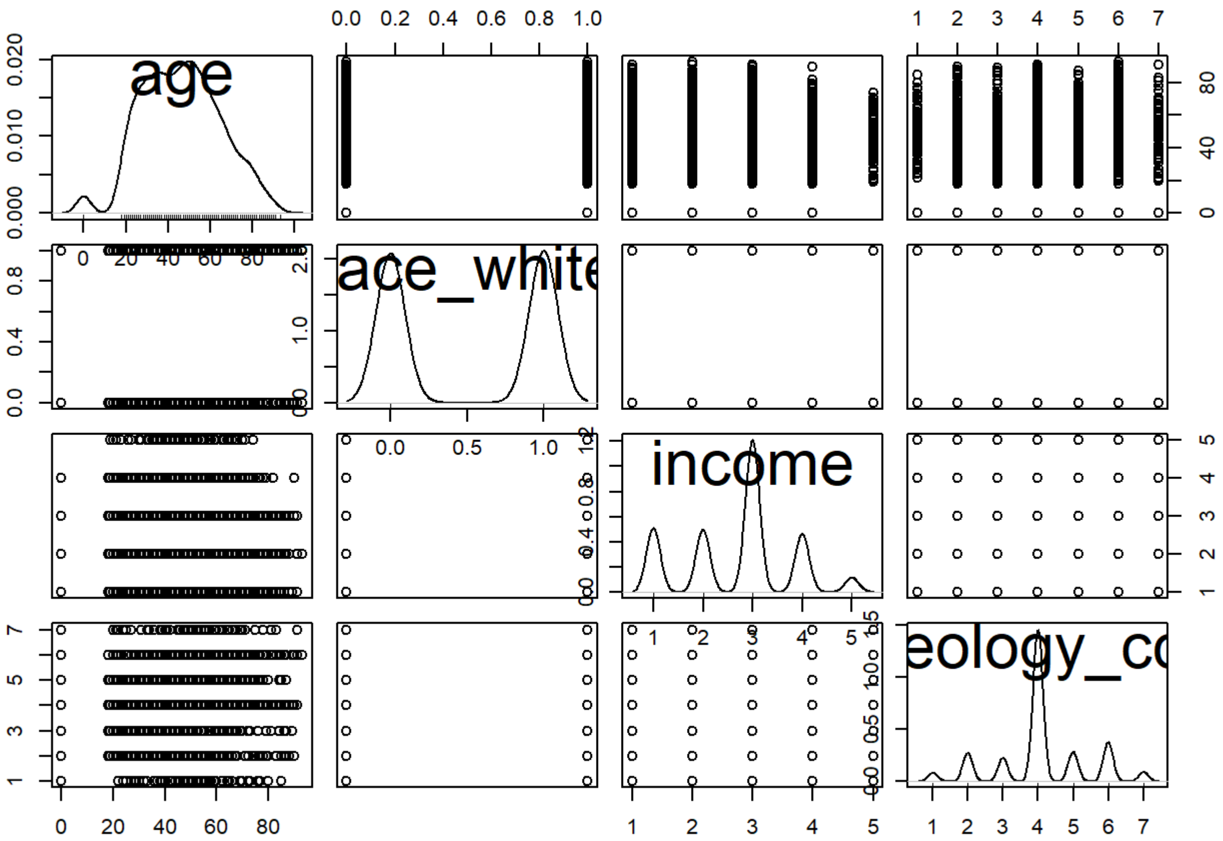
```
##
## Please cite as:
```

```
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

Quick inspection of data reveals that there are 45 records with age = 0

```
x <- data[, c("age","race_white","income","ideology_con")]
pairs(x, diag.panel=function(x){par(new=T); plot(density(x), main=''); rug(x)}, cex.labels=3)
```



```
head(table(data$age))
```

```
## 
##  0 18 19 20 21 22 
## 45 28 27 32 32 26
```

These are missing age values and we could either impute them or ignore correspnding records. Here we simply replace 0 with NA (which is R’s way of marking missing values).

```
data$age[data$age < 18] <- NA
```

```
lr <- glm(vote_rep ~ age + race_white + income + ideology_con, data=data, family="binomial")
stargazer(lr, align=TRUE, no.space=TRUE, omit.stat=c("LL","ser","f"), header=FALSE)
```

a. What’s the probability of voting Republican for a white person of average age, income, and ideology?

```
means <- sapply(lr$model, mean) # mean values for the inputs to the model.
means
```

```
##     vote_rep          age   race_white       income ideology_con 
##    0.3320158   48.6837945    0.5191041    2.8175231    4.0671937
```

```
# n <- length(means)
# white person
person <- c(intercept=1, age=NA, race_white=NA, income=NA, ideology_con=NA)
attr <- c('age', 'race_white', 'income', 'ideology_con') # person attributes
mean.p <- person
mean.p[attr] <- means[attr] # person with all attributes set to mean values
white.p <- mean.p
white.p['race_white'] <- 1 # average white person
cat('regression coefficients:\n'); print(lr$coefficients)
```

```
## regression coefficients:
```

```
##  (Intercept)          age   race_white       income ideology_con 
## -8.138839506  0.006488294  2.394180410  0.412065413  1.036101377
```

```
cat('"average" white person:\n'); print(white.p)
```

```
## "average" white person:
```

```
##    intercept          age   race_white       income ideology_con 
##     1.000000    48.683794     1.000000     2.817523     4.067194
```

Plug in our x values and use the inverse logit function to calculate probability. For binary variables, we use the modal value rather than the mean value, since it doesn’t make as much sense to talk about someone of intermediate race (for these purposes).

```
y <- sum(lr$coefficients * white.p)
p <- invlogit(y)
```

The probability that a white person votes republican, holding other variables at their mean value, is 0.48

b. What’s the change in probability of voting Republican for a person of average age, income, and ideology who switches from non-white to white?

Ans:

```
non_white.p <- mean.p
non_white.p['race_white'] <- 0
nw <- sum(lr$coefficients * non_white.p)
wh <- sum(lr$coefficients * white.p)
p.nw <- invlogit(nw)
p.wh <- invlogit(wh)
p.change <- round(p.wh - p.nw, 3)
```

c. Using the e^{β} formula from the lesson, what’s the effect on the odds ratio of shifting from black to white?

```
odds.ratio <- round(exp(lr$coefficients["race_white"]), 2)
odds.ratio
```

```
## race_white 
##      10.96
```

Shifting from non-white to white increases the odds by a factor of 10.96

d. What has a greater effect on the probability of voting Republican: an age increase of 50 years, or an incease of one income bracket? (You may choose your own baseline, such as from 25 years below average to 25 years above average; and similarly for income.)

Ans:

```
y.avg <- lr$coefficients * mean.p # prediction for the average person
y.age <- y.income <- y.avg
# Increase age term by 50 from mean.
y.age["age"] <- y.age["age"] + 50*lr$coefficients["age"]
# Increase income bracket by 1 from mean.
y.income["income"] <- y.income["income"] + 1*lr$coefficients["income"]
p.avg <- invlogit(sum(y.avg))
p.age <- invlogit(sum(y.age))
p.income <- invlogit(sum(y.income))
age.diff    <- round(p.age - p.avg, 3)    # Difference by increasing age by 50 from average
income.diff <- round(p.income - p.avg, 3) # Difference by increasing income by 1 bracket from av
erage
# cat(sprintf('age diff = %s, income diff = %s\n', signif(d.age,2), signif(d.income,2)))
```

There is a larger positive difference in the probability of voting republican by increasing the income bracket by 1 than age increase by 50 year (both from mean).

Alternatively, with odds:

```
incr.income.by.1 <- round(exp(lr$coefficients['income'] * 1),3)
incr.age.by.50 <- round(exp(lr$coefficients['age'] * 50),3)
```

i.e. there is larger odds increase of voting republican due to the increase of income bracket by 1

e.Now run the regression with all the other variables in anes_2008tr (except for voted). How do your coefficients change? What do you think explains any coefficient that became or lost significance?

```
lr_all <- glm(vote_rep ~ age + race_white + income + ideology_con + gender_male +
              education + ideology_con + partyid_rep, data=data,
              family="binomial")
stargazer(lr_all, align=TRUE, no.space=TRUE, omit.stat=c("LL","ser","f"), header=FALSE)
```

```
## SEE TABLE 2 and compare with TABLE 1
```

```
summary(lr_all)
```

```
##
## Call:
## glm(formula = vote_rep ~ age + race_white + income + ideology_con +
##     gender_male + education + ideology_con + partyid_rep, family = "binomial",
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7083  -0.3694  -0.1734   0.2452   3.2834
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.762596   0.615834 -14.229  < 2e-16 ***
## age           0.018385   0.005704   3.223  0.00127 **
## race_white    1.575573   0.204627   7.700 1.36e-14 ***
## income        0.256322   0.096789   2.648  0.00809 **
## ideology_con  0.502109   0.086146   5.829 5.59e-09 ***
## gender_male  -0.123311   0.189464  -0.651  0.51515
## education     0.012606   0.064373   0.196  0.84475
## partyid_rep   0.897118   0.057852  15.507  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1929.67  on 1517  degrees of freedom
## Residual deviance:  803.23  on 1510  degrees of freedom
##   (804 observations deleted due to missingness)
## AIC: 819.23
##
## Number of Fisher Scoring iterations: 6
```

Compare coefficients side by side:

```
all <- coef(lr_all)
sel <- coef(lr)
x <- setdiff(names(all), names(sel))
z <- rep(NA, length(x))
names(z) <- x
v <- as.data.frame(cbind(orig=c(sel, z), all))
v[['orig/all']] <- v$orig/v$all
round(v, 4)
```

```
##                 orig     all orig/all
## (Intercept)  -8.1388 -8.7626   0.9288
## age           0.0065  0.0184   0.3529
## race_white    2.3942  1.5756   1.5196
## income        0.4121  0.2563   1.6076
## ideology_con  1.0361  0.5021   2.0635
## gender_male       NA -0.1233       NA
## education         NA  0.0126       NA
## partyid_rep       NA  0.8971       NA
```

Age becomes significant when all variables are added. You can also see there are differences in the β estimates themselves. Note how party ID is also quite significant. One reason why age may now be significant where it was not before is that age both causes people to be more conservative (more likely to vote Republican), and is also correlated with being a Democrat (older people are more Democratic because being a Democrat was more popular decades ago),which of course is correlated with voting Democrat. But once you control for party ID, the second effect is controlled for, so the effect of age is no longer pulled in two directions, revealing its true effect: to make one more likely to vote Republican.