Holsenbeck_S_10

Stephen Synchronicity

2017-11-16

```
library("tidyverse")
library("dplyr")
library("htmltools")
library("forcats")
library("stargazer")
library("sjPlot")
library("car")
```

Homework Outline

14

Add at least one quadratic term into your model and interpret the results. Is it significant? What is the effect of a 1-unit increase in that variable at its mean value?

Returning to the graphs in the preliminary analysis from Homework8_9 and looking for variables that might be better explained by a quadratic term does not yield any candidates. For the purposes of answering the problem, I'll pick the % of the population with 300 min of exercise of more per week (act300), and determine whether states with moderate percentages of Republican identified individuals show higher numbers of those at this activity level and whether this percentage decreases for states with larger percentages of Republican identified individuals.

```
RObs300Fru <- lm(R ~ wgtObs + conFru + act300, Pty_Hlth)
RObs300sqFru <- lm(R ~ wgtObs + conFru + act300 + I(act300^2), Pty_Hlth)
```

```
stargazer::stargazer(RObs300sqFru, RObs300Fru, object.names = T, ci = T, single.row = T,
    star.char = c("<.1", "<.05", "<.01"), omit.table.layout = "n", keep.stat = c("aic",
        "rsq"))
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Nov 16, 2017 - 4:32:57 PM

Table 1:

	Dependent variable:	
	R	
	(1)	(2)
	RObs300sqFru	RObs300Fru
wgtObs	0.710 (−0.211, 1.632)	0.702 (−0.208, 1.613)
conFru	$0.647^{<.1}$ (−0.079, 1.372)	$0.661^{<.1}$ (−0.050, 1.372)
act300	−0.556 (−7.071, 5.960)	0.392 (−0.235, 1.019)
I(act300^2)	0.015 (−0.086, 0.115)	
Constant	−5.573 (−117.962, 106.815)	−21.001 (−59.227, 17.225)
R^2	0.312	0.311

As expected based on the graphical inspection, exploring act300 as a quadratic term did not better explain the variable's interaction with the dependent variable. The β is near 0 and did not switch sign indicating it does not decline in states with higher % of Republican identified individuals. The adjusted R^2 value has also decreased because of the addition of this quadratic term that does not fit the data.

15

Add at least one interaction term to you model and interpret the results. Is it significant? What is the effect of a 1-unit increase in one of those interacted variables holding the other at its mean value?

Based on the findings in Q11 from Homework8_9 there appears to be an interactive effect between states with higher % of individuals living a sedentary lifestyle, and not consuming fruit in one's diet and the % of individuals identifying as Republican. We will also look at a possible interaction between wgtObs & act0 due to the logical association between the two.

```
RObsFru0 <- lm(R ~ wgtObs + conFru + act0, Pty_Hlth)
RObsFru0intF0 <- lm(R ~ wgtObs + conFru + act0 + conFru * act0, Pty_Hlth)
RObsFru0intO0 <- lm(R ~ wgtObs + conFru + act0 + wgtObs * act0, Pty_Hlth)
```

```
stargazer::stargazer(RObsFru0, RObsFru0intF0, RObsFru0intO0, object.names = T, ci = T,
    single.row = T, star.char = c("<.1", "<.05", "<.01"), omit.table.layout = "n",
    keep.stat = c("aic", "rsq"))
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Nov 16, 2017 - 4:33:17 PM

Table 2:

	Dependent variable:		
	R		
	(1)	(2)	(3)
	RObsFru0	RObsFru0intF0	RObsFru0intO0
wgtObs	$0.795^{<.1}$ (−0.093, 1.683)	0.749 (−0.147, 1.646)	1.691 (−1.271, 4.653)
conFru	$0.869^{<.05}$ (0.125, 1.613)	2.093 (−0.745, 4.932)	$1.019^{<.05}$ (0.133, 1.906)
act0	$-0.713^{<.1}$ (−1.438, 0.011)	1.045 (−2.955, 5.045)	0.458 (−3.304, 4.219)
conFru:act0		−0.042 (−0.137, 0.052)	
wgtObs:act0			−0.040 (−0.166, 0.086)
Constant	−1.250 (−18.395, 15.896)	−50.017 (−160.433, 60.399)	−33.196 (−135.354, 68.961)
R^2	0.341	0.352	0.347

With the tables side-by-side the results are somewhat confusing. When the lack of fruit consumption is interacting with sedentary lifestyle, most notable is the sign shift of act0 alone from -.71 to 1.05 with a concurrent drastic change in the y-intercept from -1.25 to -50.02, the only negative values remaining in the interaction model are the interaction itself and the y-intercept. We see similar changes in the interaction between wgtObs & act0 with slight successive decreases in fit (adj R^2) with the interaction models.

The best explanation I can come up with for this phenomena is that in the reduced model without the interactions, the y-intercept indicates we start with states with a moderate % of Republicans, and states with more sedentary individuals is correlated with fewer % of the population as Republicans perhaps due to high density, urban coastal areas that are often liberal and the combination of convenience and lack of accessible outdoor recreation inclines people towards sedentary lifestyles, explaining the negative correlation.

In the model with interactive terms, the y-intercept jumps into very liberal territory at -50.02 and sedentary lifestyle has a far less significant and now positive correlation due to it's being fitted to a linear model that includes the confluence of lack of fruit consumption and sedentariness (indicating poor diet in addition to sedentary behavior) and being obese and sedentary respectively. Why the interaction terms are negative remains baffling, though the confidence intervals show that they span 0 and are statistically insignificant and thus may not be all too important. The lack of fruit being positively correlated with % of Republicans could be explained by the midwestern states where monocropped grains and factory farmed animals make up a large proportion of the agricultural product and therefore diet in contrast to places like Hawaii, California and coastal areas where fruit is more abundant.

16

Test either the model in 14 or the model in 15 using the F test for nested models. That is, estimate the full model with the variable and quadratic term, or the variable and interaction, and then estimate the reduced model without either, and run the F test to establish whether those variables significantly improve your model.

It's apparent what the result is going to be given the Adjusted R^2 values, but for the sake of practice the model with interaction between wgtObs & act0 will be tested against the reduced (original) model.

$$F = \frac{(R_c^2 - R_r^2)/df_1}{(1 - R_c^2)/df_2}$$

```
rr_org <- summary(RObsFru0)$r.squared
rc_int <- summary(RObsFru0_00)$r.squared
(f <- ((rc_int - rr_org)/1)/((1 - rc_int)/(51 - 4 - 1)))
```

```
## [1] 0.3867096
```

```
anova(RObsFru0, RObsFru0_00)
```

```
## Analysis of Variance Table
##
## Model 1: R ~ wgtObs + conFru + act0
## Model 2: R ~ wgtObs + conFru + act0 + wgtObs * act0
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     47 2352.1
## 2     46 2332.5  1    19.609 0.3867 0.5371
```

The F_{stat} and P_{value} from the F-test indicate that the interaction model was not an improvement over the original model towards a better fit for the explaining the dependent variable in a statistically significant way.

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.

```
anes <- read_csv(file = "anes_2008tr.csv")
```

```
Rep.AgeRacIncIde <- glm(vote_rep ~ age + race_white + income + ideology_con, family = "binomial",
    anes)
```

```
stargazer::stargazer(Rep.AgeRacIncIde, object.names = T, ci = T, single.row = T,
    star.char = c("<.1", "<.05", "<.01"), omit.table.layout = "n", keep.stat = c("aic",
        "rsq"))
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Nov 16, 2017 - 4:33:18 PM

Table 3:

	Dependent variable:
	vote_rep
	Rep.AgeRacIncIde
age	0.005 (−0.003, 0.013)
race_white	$2.437^{<.01}$ (2.108, 2.766)
income	$0.405^{<.01}$ (0.258, 0.551)
ideology_con	$1.042^{<.01}$ (0.912, 1.172)
Constant	$-8.077^{<.01}$ (−8.984, −7.169)
Akaike Inf. Crit.	1,180.284

a

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

I'm unsure as to whether the probability this question is asking for is best represented by the interacting terms model, or by a prediction based on the original model, so I've used both methods. Intuitively, I would guess that the prediction is what is being asked for.

```
Repint.rWhAgeIncIde <- glm(vote_rep ~ age * race_white * income * ideology_con, family = "binomial",
    anes)
```

```
stargazer::stargazer(Repint.rWhAgeIncIde, object.names = T, ci = T, single.row = T,
    star.char = c("<.1", "<.05", "<.01"), omit.table.layout = "n", keep.stat = c("aic",
        "rsq"))
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Nov 16, 2017 - 4:33:19 PM

```
input <- data.frame(race_white = 1, age = mean(anes$age), income = mean(anes$income),
    ideology_con = mean(anes$ideology_con))
(Wh <- predict.glm(Rep.AgeRacIncIde, input, type = "response"))
```

```
##         1
## 0.4860377
```

A white person of average age, income, and ideology has a probability of .49 of voting Republican, or about a 50/50 chance.

b

What's the change in probability of voting Republican for a person of average age, income, and ideology who switches from black to white?

```
input$race_white <- 0
(dp_BtoWh <- Wh - (nWh <- predict.glm(Rep.AgeRacIncIde, input, type = "response")))
```

```
##         1
## 0.4096499
```

Table 4:

	Dependent variable:
	vote_rep
	Repint.rWhAgeIncIde
age	−0.078 (−0.259, 0.102)
race_white	−4.574 (−15.578, 6.429)
income	−0.577 (−3.705, 2.552)
ideology_con	−0.636 (−2.691, 1.419)
age:race_white	0.107 (−0.108, 0.322)
age:income	0.018 (−0.047, 0.083)
race_white:income	0.897 (−2.865, 4.659)
age:ideology_con	0.020 (−0.020, 0.060)
race_white:ideology_con	1.879 (−0.638, 4.397)
income:ideology_con	0.300 (−0.414, 1.014)
age:race_white:income	−0.026 (−0.102, 0.051)
age:race_white:ideology_con	−0.024 (−0.072, 0.024)
age:income:ideology_con	−0.004 (−0.019, 0.010)
race_white:income:ideology_con	−0.265 (−1.127, 0.596)
age:race_white:income:ideology_con	0.005 (−0.012, 0.022)
Constant	−1.878 (−10.847, 7.092)
Akaike Inf. Crit.	1,168.996

The change in probability is +~.41 between a black person and a white person voting Republican when age, income and ideology are accounted for.

c

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

The change in probability:

$$\hat{P}(y=1|x=a) - \hat{P}(y=1|x=b) = \text{invlogit}(a) - \text{invlogit}(b) = \frac{e^{\beta_0+\beta_1 a_1+\beta_2 a_2+\beta_3 a_3+\beta_4 a_4}}{1+e^{\beta_0+\beta_1 a_1+\beta_2 a_2+\beta_3 a_3+\beta_4 a_4}} - \frac{e^{\beta_0+\beta_1 b_1+\beta_2 b_2+\beta_3 b_3+\beta_4 b_4}}{1+e^{\beta_0+\beta_1 b_1+\beta_2 b_2+\beta_3 b_3+\beta_4 b_4}}$$

$$\frac{e^{\beta_0+\beta_1 a}}{1+e^{\beta_0+\beta_1 a}} - \frac{e^{\beta_0+\beta_1 b}}{1+e^{\beta_0+\beta_1 b}}$$

$$\frac{e^{-8.077+0.005(46.451)+2.437(1)+0.405(2.705)+1.042(4.098)}}{1+e^{-8.077+0.005(46.451)+2.437(0)+0.405(2.705)+1.042(4.098)}} - \frac{e^{-8.077+0.005(46.451)+2.437(0)+0.405(2.705)+1.042(4.098)}}{1+e^{-8.077+0.005(46.451)+2.437(0)+0.405(2.705)+1.042(4.098)}}$$

Using R to solve that craziness...

```
b_yint <- Rep.AgeRacIncIde$coef[1]
b_age <- Rep.AgeRacIncIde$coef[2]
b_rWh <- Rep.AgeRacIncIde$coef[3]
b_inc <- Rep.AgeRacIncIde$coef[4]
b_ide <- Rep.AgeRacIncIde$coef[5]
(`1cd_BtoWh` <- (exp(b_yint + b_age * input[2] + b_rWh * 1 + b_inc * input[3] + b_ide *
    input[4])/(1 + exp(b_yint + b_age * input[2] + b_rWh * 1 + b_inc * input[3] +
    b_ide * input[4]))) - (exp(b_yint + b_age * input[2] + b_rWh * 0 + b_inc * input[3] +
```

```
    b_ide * input[4])/(1 + exp(b_yint + b_age * input[2] + b_rWh * 0 + b_inc * input[3] +
    b_ide * input[4])))
```

```
##         age
## 1 0.4096499
```

```
# % Error
as.numeric((dp_BtoWh - `1cd_BtoWh`) * 100)
```

```
## [1] 0
```

The change in odds:

$$\frac{\frac{P(y=1|x=a)}{1-P(y=1|x=a)}}{\frac{P(y=1|x=b)}{1-P(y=1|x=b)}} = \frac{e^{\beta_0}(e^{\beta_1})^{a_1}(e^{\beta_2})^{a_2}(e^{\beta_3})^{a_3}(e^{\beta_4})^{a_4}}{e^{\beta_0}(e^{\beta_1})^{b_1}(e^{\beta_2})^{b_2}(e^{\beta_3})^{b_3}(e^{\beta_4})^{b_4}} = \frac{(e^{\beta_2})^{a_2}}{(e^{\beta_2})^{b_2}} = \frac{(e^{2.437})^1}{(e^{2.437})^0}$$

Again using R to solve...

```
exp(Rep.AgeRacIncIde$coefficients)
```

```
##  (Intercept)          age   race_white       income ideology_con
## 3.107393e-04 1.004749e+00 1.143416e+01 1.498683e+00 2.834816e+00
```

```
(`1cdodds_BtoWh` <- (exp(b_rWh)^1)/(exp(b_rWh)^0))
```

```
## race_white
##   11.43416
```

The calculation by hand indicates an 11.43x higher chance of voting Republican if one is white v black given all other variables remain constant. I'm not sure how to check this answer with the predict() function.

d

What has a greater effect on the probability of voting Republican: an age increase of 50 years, or an increase 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.)

```
dage_lw <- input
dage_hi <- input
dinc_lw <- input
dinc_hi <- input
dage_lw[2] <- input[2] - 25
dage_hi[2] <- input[2] + 25
dinc_lw[3] <- input[3] - 1
dinc_hi[3] <- input[3] + 1
(`50yrs` <- predict.glm(Repint.rWhAgeIncIde, dage_hi, type = "response") - predict.glm(Repint.rWhAgeInc
    dage_lw, type = "response"))
```

```
##          1
## 0.03083505
```

```
(inc_1 <- predict.glm(Repint.rWhAgeIncIde, dinc_hi, type = "response") - predict.glm(Repint.rWhAgeIncIde
    dinc_lw, type = "response"))
```

```
##        1
## 0.116571
```

```
inc_1/`50yrs`
```

```
##       1
## 3.78047
```

If you are of non-white race, the change in the probability of voting Republican from an age of 21.5 to 71.5 is +~3%, whereas the the change in probability of voting Republican from one income bracket below average to one income bracket above average is +~11%. In other words, a non-white person is ~3.78 times more likely to vote Republican by moving up two income brackets than by aging 50 years.

```
dage_lw[1] <- 1
dage_hi[1] <- 1
dinc_lw[1] <- 1
dinc_hi[1] <- 1
(`50yrswh` <- predict.glm(Repint.rWhAgeIncIde, dage_hi, type = "response") - predict.glm(Repint.rWhAgeIn
    dage_lw, type = "response"))
```

```
##          1
## 0.01946875
```

```
(inc_1wh <- predict.glm(Repint.rWhAgeIncIde, dinc_hi, type = "response") - predict.glm(Repint.rWhAgeInc
    dinc_lw, type = "response"))
```

```
##         1
## 0.1293148
```

```
inc_1wh/`50yrswh`
```

```
##        1
## 6.642172
```

If you are white, the change in the probability of voting Republican from an age of 21.5 to 71.5 is +~2%, whereas the the change in probability of voting Republican from one income bracket below average to one income bracket above average is +~13%. In other words, a white person is ~6.64 times more likely to vote Republican by moving up two income brackets than by aging 50 years.

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?

```
Rep.AgeRacGenEduIncIdePid <- glm(vote_rep ~ age + race_white + gender_male + education +
    income + ideology_con + partyid_rep, family = "binomial", anes)
```

```
stargazer::stargazer(Rep.AgeRacGenEduIncIdePid, Rep.AgeRacIncIde, object.names = T,
    ci = T, single.row = T, star.char = c("<.1", "<.05", "<.01"), omit.table.layout = "n",
    keep.stat = c("aic", "rsq"))
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Nov 16, 2017 - 4:33:20 PM

Table 5:

	Dependent variable:	
	vote_rep	
	(1)	(2)
	Rep.AgeRacGenEduIncIdePid	Rep.AgeRacIncIde
age	$0.015^{<.01}$ (0.005, 0.025)	0.005 (−0.003, 0.013)
race_white	$1.627^{<.01}$ (1.228, 2.025)	$2.437^{<.01}$ (2.108, 2.766)
gender_male	−0.140 (−0.508, 0.229)	
education	0.019 (−0.106, 0.145)	
income	$0.250^{<.01}$ (0.061, 0.439)	$0.405^{<.01}$ (0.258, 0.551)
ideology_con	$0.511^{<.01}$ (0.343, 0.679)	$1.042^{<.01}$ (0.912, 1.172)
partyid_rep	$0.894^{<.01}$ (0.782, 1.007)	
Constant	$-8.646^{<.01}$ (−9.829, −7.463)	$-8.077^{<.01}$ (−8.984, −7.169)
Akaike Inf. Crit.	830.593	1,180.284

```
sumCom <- summary(Rep.AgeRacGenEduIncIdePid)
```

Considering each factor individually:

age

Age is a significant factor in the model that includes all available variables and insignificant in the reduced model, indicating that age plays a significant role in voting Republican when gender, education and party_id are factored in. However, it's influence accounts for only a 2% change in voting Republican with each additional year.

race_white

Race is by and large the most significant and influential predictive variable in voting Republican. In the reduced model, it accounts for ~11.4x increase in the chance of voting Republican. When gender, education, and party_id are included the influence of the variable is mitigated to increasing the odds by ~5.1x. This could be due to a suppression effect with the addition of the liberalizing factor, education (liberalizing for all but those whom attained a high school degree and decided not to continue their education), or a sharing of influence with party_id where higher values also indicate conservatism and is therefore also closely linked to voting Republican.

gender_male

The addition of gender has a insignificant slightly negative effect on the probability of voting Republican in the model, this is likely due to a suppression effect in a model with significantly stronger predictors.

education

Education has a insignificant albeit slightly positive effect on P(vote_rep), perhaps due to it's inclusion in a linear model when we know from previous exploration that it has a quadratic effect.

income

Income is a stronger, highly significant influence in the reduced model, and retains it's significance though with a more muted effect in the complete model. This could be due to multicollinearity between education and income, given we know education has a quadratic effect. In other words, given those who have more education are likely to have greater income, and are also less likely to vote Republican, the effect of income would be suppressed in a model where age is statistically significant and that includes education.

ideology_con

A conservative ideology is going to exhibit colinearity with the response (dependent) variable which we can observe reflected in its significance level in both models. It's effect is mitigated in the complete model, possibly due to its chained causation with the stronger indicator of party_id.

partyid_rep

Being affiliated with the Republican party is going to have the strongest colinearity with voting Republican, reflected in it's significance and placeholder as the 2nd highest coefficient. It's magnitude of effect is second to that of being white. However, this variable has a wide-range of values which are minimally explained:

partyid_rep is the party ID of the respondent, with Republican being the higher value

I could not find in the ANES documentation what variable and or question on the survey this was matched to, which would help to better explain it.