Puzzle Solving Case Study:

Latinos and Discrimination

Sergio I Garcia-Rios

Table of Contents

## Preliminary set up

Let’s load the data. We are going to use the Latino National Survey for this workshop. This particular version of the LNS comes in .dta format which is the extension used for a statistical package called Stata

Stata is very popular and powerful and many statisticians and social scientist use it. I used to swear by it. But then, I met R and… well things change.

The good thing is that R allows you to read in many different files, yes there’s a pack for that!

### Load Libraries/Packages

We are going to use the following packages, make sure you have them installed

library(haven)  
library(labelled)  
library(tidyverse)  
library(pander)  
library(forcats)  
library(stargazer)

### Load Data

OK now let’s really lead the data, I use read\_dta here but depending on the extension of your file that will change.

lns <- read\_dta("lns\_full.dta")

This is going to get somewhat technical but just trust me here… for this Workshop we are going to do some data transformation, if your data came as a Stata file sometimes is good to run this command so you can convert all files into factors. You can do that also just for a single variable (I actually recommend this) when you need to graph or something like that.

lns <- haven::as\_factor(lns, only\_labelled = TRUE)

## Analysis

Now we are ready to start our analysis. Recently a a very interesting came out, [here](https://www.dol.gov/_sec/media/reports/hispaniclaborforce/) and shows that while Latinos face do face higher rates of unemployment they seem to be significantly more pessimistic about their economic outlook than other Whites and Blacks. That’s sounds interesting and I think it deserves further exploration.

### Main Issue

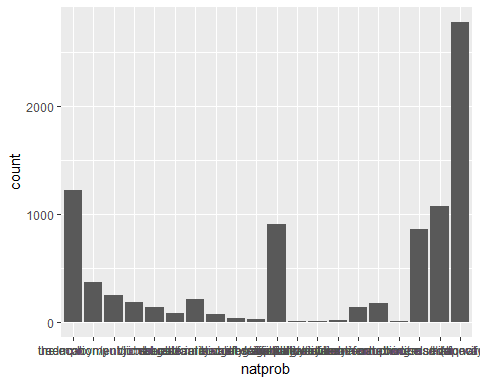
Let’s begin by looking at what Latinos think is the main issue facing the country. The variable is natprob

lns %>% count(natprob) %>%   
 mutate(pct = prop.table(n)) %>% pander()

|  |  |  |
| --- | --- | --- |
| natprob | n | pct |
| the economy | 1222 | 0.1415 |
| unemployment/jobs | 377 | 0.04366 |
| education/public schools | 250 | 0.02896 |
| crime | 187 | 0.02166 |
| drugs | 146 | 0.01691 |
| health care | 82 | 0.009497 |
| race relations | 213 | 0.02467 |
| values/family values/morality | 77 | 0.008918 |
| budget deficit | 37 | 0.004285 |
| social security/care for the elderly | 32 | 0.003706 |
| illegal immigration | 907 | 0.105 |
| affirmative action | 9 | 0.001042 |
| welfare/welfare reform | 15 | 0.001737 |
| environment | 23 | 0.002664 |
| political system/corruption/scandal | 140 | 0.01621 |
| foreign policy/international concerns/national defense | 180 | 0.02085 |
| abortion | 11 | 0.001274 |
| something else (specify) | 865 | 0.1002 |
| dk/rf | 1079 | 0.125 |
| iraq war | 2782 | 0.3222 |

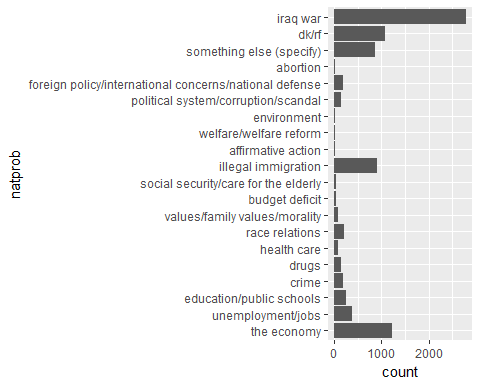
Seems like the Iraq war is the main issue at the bottom pf the table. Let’s try to plot it with a bar-graph

ggplot(lns, aes(x = natprob)) +  
 geom\_bar()



WOW… That looks pretty bad, let’s try to improve it by flipping the axes with coord\_flip

ggplot(lns, aes(x = natprob)) +  
 geom\_bar() +   
 coord\_flip()



That’s better but we can make it much better if we do some recoding to the natprob variable.

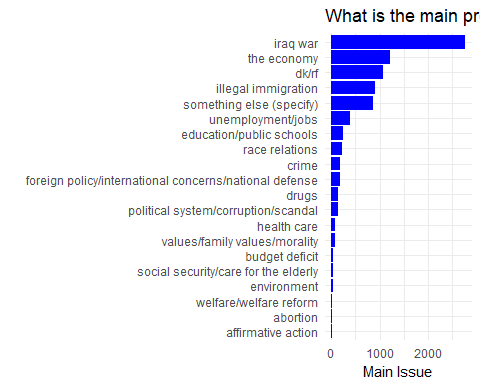
* First we create a new variable called natprob\_r
* This variable is going to be exactly the same as natprob but ordered by frequency. We do that with fct\_infreq
* Then, we reverse the order with fct\_rev so that it goes from high to low

lns <- lns %>% mutate(natprob\_r =   
 haven::as\_factor(natprob) %>%   
 fct\_infreq() %>%   
 fct\_rev())

Now we can create the graph, notice that I am adding some other elements:

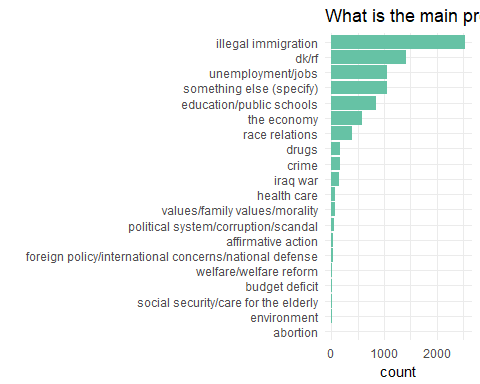
* I added another aesthetic: fill and specified that I wanted the bar to be filled with the color blue
* I also added a new theme. I am at point in my life wheretheme\_minimal is my favorite but I also had stage in my life where I liked theme\_bw… things change…

ggplot(lns, aes(x = natprob %>%   
 fct\_infreq() %>%   
 fct\_rev()  
 )) +   
 geom\_bar( fill = "blue") + # added blue as the color for the bars  
 coord\_flip() + # still wnat teh coordinted flipped  
 labs(x = "",  
 y = "Main Issue",  
 title= "What is the main problem facing the country") + # here I can specify labels and titles   
 theme\_minimal()



What about the issues for Latinos? Let’s do the same but now, I actually don’t really Like that blue color, so let’s borrow a nice green color from [ColorBrewer](http://colorbrewer2.org)

ggplot(lns, aes(x =   
 latprob %>%   
 fct\_infreq %>%   
 fct\_rev())) +  
 geom\_bar(fill = "#66c2a5") +  
 coord\_flip() +  
 labs(x = "",   
 title= "What is the main problem facing Latinos") +  
 theme\_minimal()



### Poor people can get ahead if they work hard

Yes, as we expected, Latinos care a lot about the election but also about unemployment. Let’s now really see whether Latinos are pessimistic about their economic outlook. The poordisc variable asks the following question:

* How strongly do you agree or disagree with the following: *Poor people can get ahead in life if they work hard*

Let’s explore that variable

prop.table(table(lns$poordisc))

##   
## strongly disagree somewhat disagree somewhat agree strongly agree   
## 0.033 0.033 0.183 0.722   
## dk   
## 0.028

Well seems like most Latinos agree that working hard will get them ahead in life but let’s see more in depth because Latinos are not a monolithic group.

Let’s look at them by generation. I created this variable newgen that breaks down Latinos by generation:

prop.table(table(lns$newgen))

##   
## 1 1.5 2 2.5 3 4   
## 0.642 0.067 0.118 0.029 0.066 0.078

Now, lets look at both. That is, what Latinos think about their economic chances if they work hard by generation. Notice I am adding a 2 at the end, this makes the prop.table show percentages by column.

table(lns$poordisc, lns$newgen)

##   
## 1 1.5 2 2.5 3 4  
## strongly disagree 135 20 51 8 32 37  
## somewhat disagree 121 13 46 19 35 51  
## somewhat agree 879 103 209 57 145 185  
## strongly agree 4227 422 674 161 343 384  
## dk 159 17 34 8 14 12

prop.table(table(lns$poordisc,   
 lns$newgen), 2)

##   
## 1 1.5 2 2.5 3 4  
## strongly disagree 0.024 0.035 0.050 0.032 0.056 0.055  
## somewhat disagree 0.022 0.023 0.045 0.075 0.062 0.076  
## somewhat agree 0.159 0.179 0.206 0.225 0.255 0.277  
## strongly agree 0.766 0.734 0.665 0.636 0.603 0.574  
## dk 0.029 0.030 0.034 0.032 0.025 0.018

You can see that the percentage of those strongly agreeing drops from 77 to 57 as you go across generations. Also, notice that dk… it means don’t know, we don’t need it so let’s drop it using recode

lns <- lns %>%   
 mutate(poordisc\_r =   
 recode(poordisc,   
 "dk" = NA\_character\_))

Let’s visualize these data, but first we need to do some recoding to the variable to make sure that the values appear in the correct order, usually they do but in case they don’t I want to show you how to specify the order.

* So, again, we go the dataframe “lns”,
* **then** with mutate create a variable called poor\_r,
* **then** specify the order with re\_level.

lns <- lns %>% mutate(poord\_r =   
 fct\_relevel(poordisc\_r, "strongly disagree", "somewhat disagree",   
 "somewhat agree", "strongly agree"))

Now we create a proportion table like we did with proptable but now we put it in a data frame called poor\_can.

poor\_can<- lns %>%  
 group\_by(newgen) %>%   
 count(poord\_r) %>%  
 mutate(prop = prop.table(n)) %>%   
 na.omit()

This is what we are doing step by step:

* In an new object called poor\_can we
* go to the lns
* **then** we generate counts for newgen and pood\_r
* **then** with mutate we create a column called

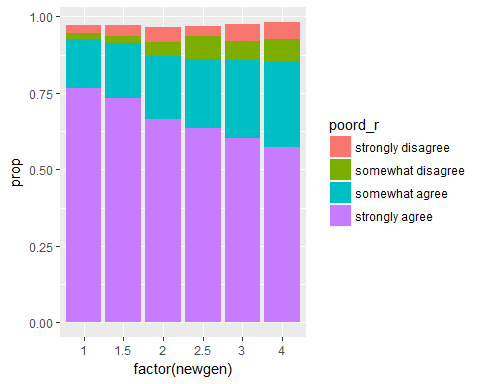
And this is what it looks like:

poor\_can

## # A tibble: 24 x 4  
## # Groups: newgen [6]  
## newgen poord\_r n prop  
## <dbl> <fctr> <int> <dbl>  
## 1 1.0 strongly disagree 135 0.024  
## 2 1.0 somewhat disagree 121 0.022  
## 3 1.0 somewhat agree 879 0.159  
## 4 1.0 strongly agree 4227 0.766  
## 5 1.5 strongly disagree 20 0.035  
## 6 1.5 somewhat disagree 13 0.023  
## 7 1.5 somewhat agree 103 0.179  
## 8 1.5 strongly agree 422 0.734  
## 9 2.0 strongly disagree 51 0.050  
## 10 2.0 somewhat disagree 46 0.045  
## # ... with 14 more rows

Now we are ready to create a graph using those proportions, that is, the poor\_can mini data set that we created.

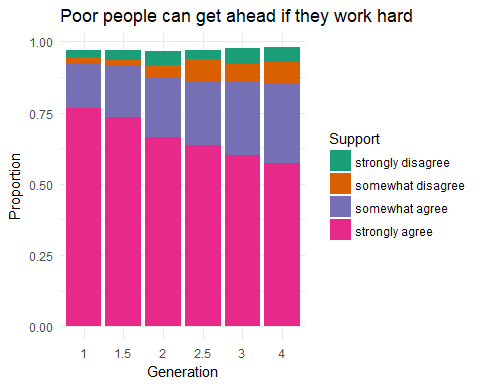
ggplot(poor\_can, aes(x=factor(newgen), y = prop, fill = poord\_r)) +  
 geom\_bar(stat= "identity")



Here is the same graph with some improvements, I added:

* Labels and titles with labs
* A different color palette with scale\_fill\_brewer I am using the palette Dark2 from ColorBrewer
* A nice and clean predefined theme called theme\_minimal

ggplot(poor\_can, aes(x = factor(newgen), y = prop,  
 fill = poord\_r)) +  
 geom\_bar(stat= "identity") +  
 labs(x = "Generation",   
 y = "Proportion",  
 title= "Poor people can get ahead if they work hard",  
 fill = "Support") +   
 scale\_fill\_brewer(palette = "Dark2") +  
 theme\_minimal()



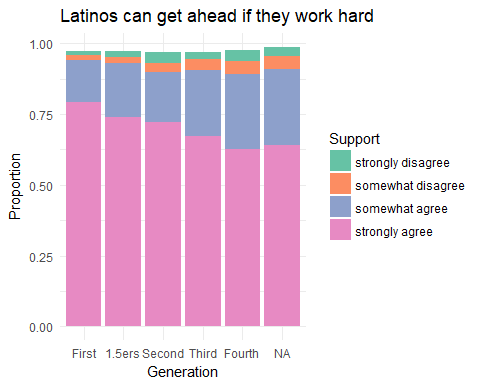
### Latinos can get ahead if they work hard

They also asked a similar question but related to Latinos specifically:

* How strongly do you agree or disagree with the following: *Latinos can get ahead in life if they work hard*

The name of he variable is latdic. We are going to do the same graph but now with this variable, notice that I am adding a new line to specify labels on the x axis with scale\_x\_discrete (those are called axis ticks, by the way)

lns <- lns %>% mutate(latdisc\_r = recode(latdisc, 'dk' = NA\_character\_))   
  
  
latinos\_can<- lns %>%  
 group\_by(newgen) %>%   
 count(latdisc\_r) %>%  
 mutate(prop = prop.table(n)) %>%   
 na.omit()  
  
  
  
ggplot(latinos\_can, aes(x = factor(newgen), y = prop,  
 fill = latdisc\_r)) +  
 geom\_bar(stat= "identity") +  
 theme\_minimal() +  
 labs(x = "Generation",   
 y = "Proportion",  
 title= "Latinos can get ahead if they work hard",  
 fill = "Support") +   
 scale\_fill\_brewer(palette = "Set2") +  
 scale\_x\_discrete(labels = c("First", "1.5ers", "Second", "Third", "Fourth"))



### Discrimantion

So Latinos think that it is hard to get ahead in life not only for poor people but for Latinos in particular.

I want to investigate this further, and I think it has to do with perceptions of discrimination. The LNS asked Latinos four questions to see if they have experienced some kind of discrimination. These are the questions:

* Have you ever … been unfairly fired or denied a job or promotion?
* Have you ever … been unfairly treated by the police?
* Have you ever … been unfairly prevented from moving into a neighborhood (vecindario o barrio) because the landlord or a realtor refused to sell or rent you a house or apartment?
* Have you ever … been treated unfairly or badly at restaurants or stores?

IF they said yes to **any** of those, they followed up asking why they think they were discriminated, the variable is whydisc

lns %>% count(whydisc)

## # A tibble: 10 x 2  
## whydisc n  
## <fctr> <int>  
## 1 being latino 856  
## 2 being an immigrant 232  
## 3 your national origin 220  
## 4 your language or accent 378  
## 5 your skin color 362  
## 6 your gender 66  
## 7 your age 105  
## 8 other 438  
## 9 dk/na 211  
## 10 <NA> 5766

Seems like simply being Latino is the main reason why Latinos report they have been discriminated.

Now let’s look at that across generations, again we begin by recoding that dk/na with mutate and Recode

lns<- lns %>% mutate(whydisc\_r = recode(whydisc, "dk/na" = NA\_character\_))

Now look at the proportion table

prop.table(table(lns$whydisc\_r,lns$newgen), 2)

##   
## 1 1.5 2 2.5 3 4  
## being latino 0.3206 0.3756 0.3517 0.3394 0.2913 0.2747  
## being an immigrant 0.1487 0.0732 0.0287 0.0092 0.0000 0.0123  
## your national origin 0.0605 0.1122 0.1077 0.0826 0.1299 0.0895  
## your language or accent 0.2242 0.0976 0.0742 0.0734 0.0354 0.0309  
## your skin color 0.0755 0.1366 0.1770 0.2202 0.2126 0.2438  
## your gender 0.0172 0.0195 0.0287 0.0275 0.0354 0.0432  
## your age 0.0269 0.0488 0.0359 0.0550 0.0669 0.0648  
## other 0.1263 0.1366 0.1962 0.1927 0.2283 0.2407

You can see interesting patterns:

* For example, being Latino seems to be salient across all generations but
* Accent and being immigrant is, as expected only salient among first generations
* But while later generations will not have access and wouldn’t be immigrants they still feel discriminated an now seem to attribute it to their skin color

Now I actually want to go back to those individual questions about acts of discrimination. I want to see whether people have experienced those more than once, so I created an Index, each of those variables has a 1 if they said yes or 0 if they said no. So basically I just added all the variables (of course I first recoded them to get rid of the dk/na).

For example if someone said yes to

* unfairly fired or denied a job or promotion` **and**
* also said yes to been treated unfairly or badly at restaurants or stores
* they would have “collected” two 1’s so ending up with a discrimination score of 2

This is how you do the index:

lns <- lns %>% mutate(discindex = dfired\_r + dbadpol\_r+ dhousing\_r+drestaur\_r)

Let’s take a look:

prop.table(table(lns$discindx))

##   
## 1 2 3 4   
## 0.774 0.179 0.024 0.023

How does that look across generations (notice how here I am multiplying by 100, to make it look like percentages)

prop.table(table(lns$discindx, lns$newgen), 2)\*100

##   
## 1 1.5 2 2.5 3 4  
## 1 81.3 76.0 74.3 69.4 64.3 64.9  
## 2 15.3 19.5 18.4 24.1 27.2 27.2  
## 3 1.8 2.1 3.3 4.1 4.7 4.5  
## 4 1.6 2.3 4.1 2.4 3.8 3.3

## Regression and predicted probabilities

I am going to run now a regression usingpoord\_r as my dependent variable. To do that I first have to tell R that this time the variable poord\_r will be numeric

lns$poord\_r <- as.numeric(lns$poord\_r)  
  
  
summary(model1<-lm(poord\_r ~ newgen, data = lns))

##   
## Call:  
## lm(formula = poord\_r ~ newgen, data = lns)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.709 -0.350 0.291 0.291 0.650   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.82822 0.01502 254.8 <2e-16 \*\*\*  
## newgen -0.11952 0.00827 -14.4 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7 on 8355 degrees of freedom  
## (277 observations deleted due to missingness)  
## Multiple R-squared: 0.0244, Adjusted R-squared: 0.0243   
## F-statistic: 209 on 1 and 8355 DF, p-value: <2e-16

What this shows is that there is a very strong **negative** relationship between feeling like working hard is not enough and generation.

The estimate -0.11952, so based on these data every subsequent generation would be around -0.11952 lower than the previous one.

We can use predict to extract the predicted probabilities.

-First we create a hypothetical case. -In this example I just want to look at First vs Second generations

hypdata<- data.frame(newgen=c(1,2))

Here is the command. The result show that the predicted level (in a 1-4 scale) of saying that poor people can get ahead by working hard drops from 3.7 to 3.6 when you compare First vs Second

predict(model1, hypdata)

## 1 2   
## 3.7 3.6

I can also look at more than two, what about First, 1.5ers and Fourth

hypdata<- data.frame(newgen=c(1, 1.5, 4))  
predict(model1, hypdata)

## 1 2 3   
## 3.7 3.6 3.4

Now I am going to run a full model including the following variables:

* Generation
* Our discrimination index
* Level of education
* Age
* Whether the respondent is Mexican
* Income
* whether the respondent is female

summary(fullmodel <- lm(poord\_r ~ newgen + discindx + trieduc + age + mexican + income + female, data = lns))

##   
## Call:  
## lm(formula = poord\_r ~ newgen + discindx + trieduc + age + mexican +   
## income + female, data = lns)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.9324 0.0105 0.1067 0.1514 2.4636   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.667033 0.032093 145.42 <2e-16 \*\*\*  
## newgen -0.044805 0.007541 -5.94 3e-09 \*\*\*  
## discindx -0.751601 0.010335 -72.73 <2e-16 \*\*\*  
## trieduc -0.013688 0.008851 -1.55 0.122   
## age 0.000809 0.000440 1.84 0.066 .   
## mexican 0.038185 0.014388 2.65 0.008 \*\*   
## income -0.001276 0.003669 -0.35 0.728   
## female -0.016312 0.013128 -1.24 0.214   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.52 on 6414 degrees of freedom  
## (2212 observations deleted due to missingness)  
## Multiple R-squared: 0.468, Adjusted R-squared: 0.467   
## F-statistic: 806 on 7 and 6414 DF, p-value: <2e-16

Now for my hypothetical values I am only going to move generation and whether the respondent is Mexican the rest I am going to hold them at their mean

hypdata\_mex<- lns %>% with(expand.grid(newgen=c(1,4),  
 discindx = mean(discindx, na.rm = TRUE),   
 trieduc = mean(trieduc, na.rm = TRUE),   
 age = mean(age, na.rm = TRUE),   
 mexican = c(0,1),  
 income = mean(income, na.rm = TRUE),   
 female = mean(female, na.rm = TRUE)))

This is what out hypothetical data looks like, one for each generation requested (1, 4) and one for each nationality (Mexican/non-Mexican)

hypdata\_mex

## newgen discindx trieduc age mexican income female  
## 1 1 1.3 2 41 0 3.5 0.55  
## 2 4 1.3 2 41 0 3.5 0.55  
## 3 1 1.3 2 41 1 3.5 0.55  
## 4 4 1.3 2 41 1 3.5 0.55

Now I can get predicted probabilities for this full model with those two variables “moving”. I am also adding some confidence interval

predict\_df\_mex <- predict(fullmodel, hypdata\_mex, interval = "confidence", level = .90)

Let’s look at the result:

predict\_df\_mex

## fit lwr upr  
## 1 3.6 3.6 3.7  
## 2 3.5 3.5 3.5  
## 3 3.7 3.7 3.7  
## 4 3.5 3.5 3.6

We have four columns,

* fit This is the estimated value
* lwr This is the lower bound in our confidence interval
* upr This is the upper bound in our confidence interval

And we have four rows, one for each hypothetical value that we wanted. To facilitate graphing I am going to add the hypdata\_mex to the predict\_df\_mex using the command cbind

plot\_predicts\_mex <- cbind(predict\_df\_mex, hypdata\_mex)

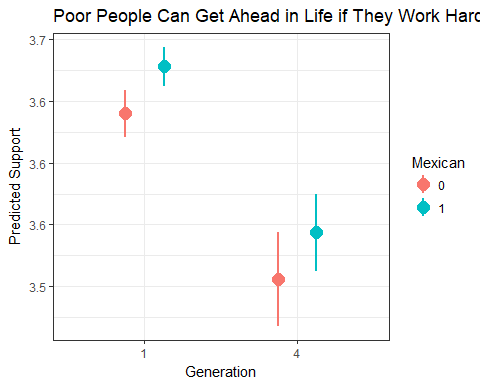
It looks like this now

plot\_predicts\_mex

## fit lwr upr newgen discindx trieduc age mexican income female  
## 1 3.6 3.6 3.7 1 1.3 2 41 0 3.5 0.55  
## 2 3.5 3.5 3.5 4 1.3 2 41 0 3.5 0.55  
## 3 3.7 3.7 3.7 1 1.3 2 41 1 3.5 0.55  
## 4 3.5 3.5 3.6 4 1.3 2 41 1 3.5 0.55

With that mini dataset that I created I can make a nice graph

ggplot(plot\_predicts\_mex, aes(y = fit, x = factor(newgen),  
 color = factor(mexican),  
 ymin = lwr,  
 ymax= upr)) +  
 geom\_pointrange(size = 1, position = position\_dodge(width = .5)) +  
 theme\_bw() +  
 labs(x = "Generation",   
 y = "Predicted Support",  
 title = "Poor People Can Get Ahead in Life if They Work Hard by Country of Origin",  
 color = "Mexican")

 Now let’s look at discrimination. That’s the only variable I am going to “move” for this graph, the rest will be held at their mean.

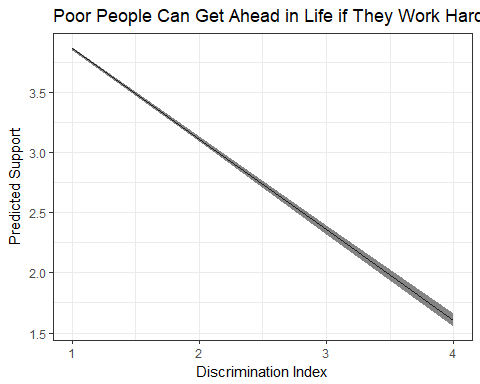
hypdata2<- lns %>% with(expand.grid(newgen= mean(newgen, na.rm = TRUE),  
 discindx = seq(1,4),   
 trieduc = mean(trieduc, na.rm = TRUE),   
 age = mean(age, na.rm = TRUE),   
 mexican = mean(mexican, na.rm = TRUE),  
 income = mean(income, na.rm = TRUE),   
 female = mean(female, na.rm = TRUE)))  
  
predict\_df2 <- predict(fullmodel, hypdata2, interval = "confidence")

Now extract predicted probabilities

plot\_predicts2 <- cbind(predict\_df2, hypdata2)

And graph

ggplot(plot\_predicts2, aes(x = discindx, y = fit,  
 ymin = lwr,  
 ymax= upr)) +  
 geom\_line() +  
 geom\_ribbon(alpha = .6) +  
 theme\_bw() +  
 labs(x = "Discrimination Index",   
 y = "Predicted Support",  
 title = "Poor People Can Get Ahead in Life if They Work Hard")



Yes… The more that Latinos experienced discrimination the less they think that it is possible to get ahead in life with hard work. It might be the case that the experience and perceptions of discrimination is also passed down by generation, making further generation more pessimistic.