Data Visualization class

Limor Raviv

8 November 2018

Intoduction

The following Rmarkdown document includes a detailed example of how to visualize and analyze data, plot regression models, calculate p-values and more.

We'll use the "GSSvocab" dataset, which contains information from the General Social Survey (GSS) of the University of Chicago. It inlcudes vocabulary scores collected over the course of 20 years from over 28,000 people. We'll analyze the vocabulary scores by individuals' age, gender, education level and nativeness.

Feel free to reuse and edit any part of this document/code!

What's Rmarkdown?

This is an R Markdown document. It basically combines text with R code (models, plots etc), and can be used to create beautiful HTMLs, PDFs, Word documents, slides and even websites.

When you click on the "Knit" button on top, it will generate a document that includes all the specified content: this text, as well as the output of any embedded R code chunks within the document (unless you decide not to include it in your final output).

For more details on using R Markdown see http://rmarkdown.rstudio.com.

I also recommend using this awesome cheat sheet: https://github.com/rstudio/cheatsheets/raw/master/rmarkdown-2.0.pdf

Text in Rmarkdown

Text appears like this, on a white background.

You can format the text to be **bold** or *italics*, and have it appear in different sizes by starting a line with hashtags:

for main headers

for subheaders

for subsubheaders

You can also use lists and bullet-points:

1. This
2. is
3. a list

- and
- these

- are
- bullets

And even write nice equations by using the dollar sign:

$$\frac{n!}{k!(n-k)!} = \binom{n}{k}$$

Code in Rmarkdown

Chuncks of code appear in a grey box denoted by "‘ at the begining and end, and have a curly-brackets header (see below). The code always starts with some name, and then some technical instructions (e.g., do you want to include the actual code in the document or just the output? Do you want to see warnings?). For example "echo=TRUE" means I want the code itself to appear in the final file (not just an output, if any). Check out the cheat sheet for more details.

```
example <- 1.987
```

You can also include some R code inside the text by using your code in grave accents. For exmaple, 2 multiplied by 10 equals 20. This can be used to integrate values from your enviroment (like beta-coefficients) in the actual text without the need to copy them, like the value from the exmaple above is 1.987.

Let's get started!

For editing and running the code, please install and load the following packages first.

Note that "include=FALSE" here means that this chunck of code will not appear in the final document.

The dataset

Now, let's load the dataset and play with it a bit to see what's going on.

```
##       year          gender      nativeBorn    ageGroup        educGroup
##  1994   : 1977   female:16385   no  : 2556   18-29:5849   <12 yrs  :5924
##  1996   : 1960   male  :12482   yes :26224   30-39:6248   12 yrs   :8612
##  2016   : 1888                  NA's:   87   40-49:5246   13-15 yrs:7182
##  1982   : 1860                               50-59:4329   16 yrs   :3914
##  1987   : 1819                               60+  :7101   >16 yrs  :3154
##  2014   : 1675                               NA's :  94   NA's     :  81
##  (Other):17688
##      vocab             age             educ
##  Min.   : 0.000   Min.   :18.00   Min.   : 0.00
##  1st Qu.: 5.000   1st Qu.:32.00   1st Qu.:12.00
##  Median : 6.000   Median :44.00   Median :12.00
##  Mean   : 5.998   Mean   :46.18   Mean   :13.04
##  3rd Qu.: 7.000   3rd Qu.:59.00   3rd Qu.:15.00
##  Max.   :10.000   Max.   :89.00   Max.   :20.00
##  NA's   :1348     NA's   :94      NA's   :81
```

```
##
##              female  male
##   Non-native   1398  1158
##   Native      14936 11288
```

```
##
##         female male
##   18-29   3214 2635
##   30-39   3592 2656
##   40-49   2838 2408
##   50-59   2403 1926
##   60+     4275 2826
```

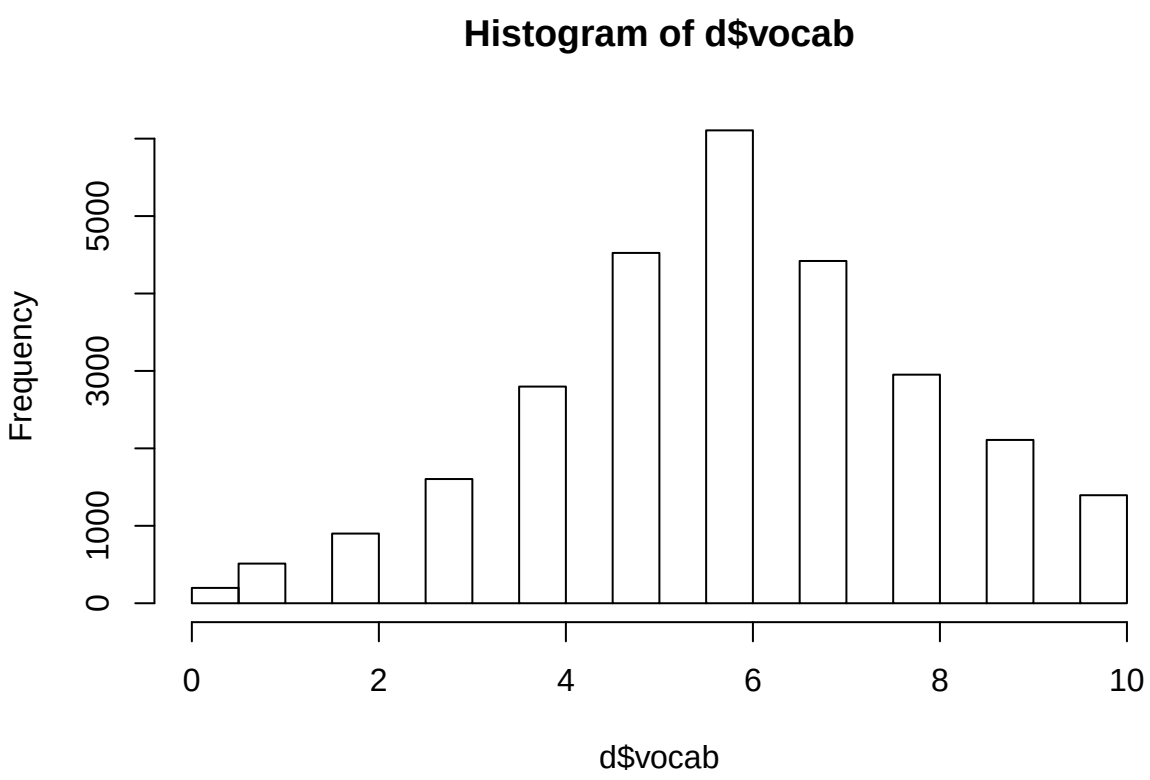
```
##
##             female male
##   <12 yrs     3415 2509
##   12 yrs      5094 3518
##   13-15 yrs   4156 3026
##   16 yrs      2103 1811
##   >16 yrs     1574 1580
```

```
##
##         Non-native Native
##   18-29        497   5343
##   30-39        674   5551
##   40-49        523   4706
##   50-59        390   3931
##   60+          463   6614
```

```
##
##             Non-native Native
##   <12 yrs          690   5215
##   12 yrs           520   8067
##   13-15 yrs        586   6583
##   16 yrs           369   3538
##   >16 yrs          381   2765
```

```
##
##          Non-native Native
##   female       1398  14936
##   male         1158  11288
```

```
##
##         <12 yrs 12 yrs 13-15 yrs 16 yrs >16 yrs
##   18-29    1063   1848      1788    775     368
##   30-39     812   1785      1779   1024     833
##   40-49     790   1528      1345    835     737
##   50-59     881   1306      1002    576     554
##   60+      2365   2118      1240    693     655
```



Predictions?

Before we ran any models and look at the data, what are your predictions? What do you think will affect vocabulary scores?

For each of these variables, write down: Do you think this variable matters? How so? Should it be a fixed effect or a random effect?

- year (of data collection)
- gender
- nativeBorn (in the USA)
- age
- educ (years of formal education)

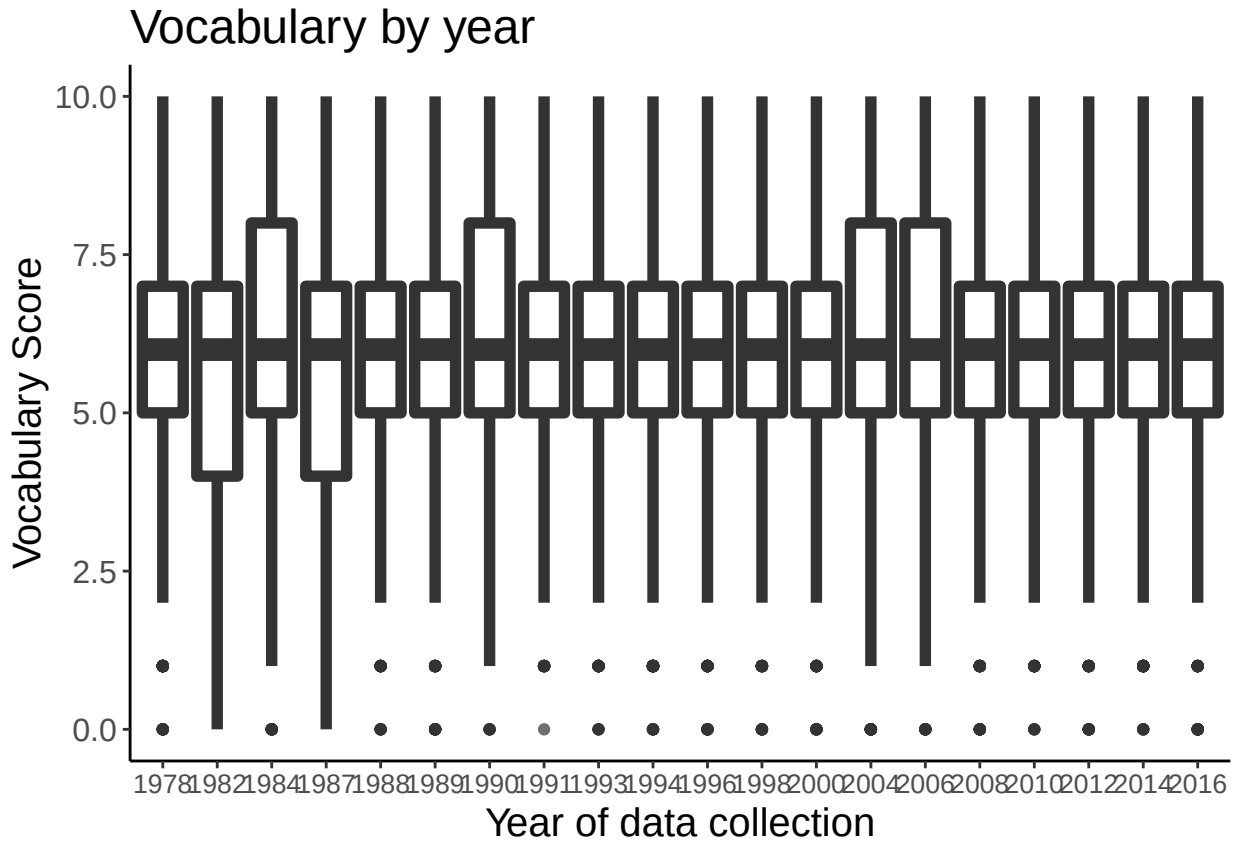
Think: Do you expect any interactions? For example, it's possible that being native to the US will moderate the effect of education and/or age. DO you think this relation will be further moderated by gender?

Plotting the data

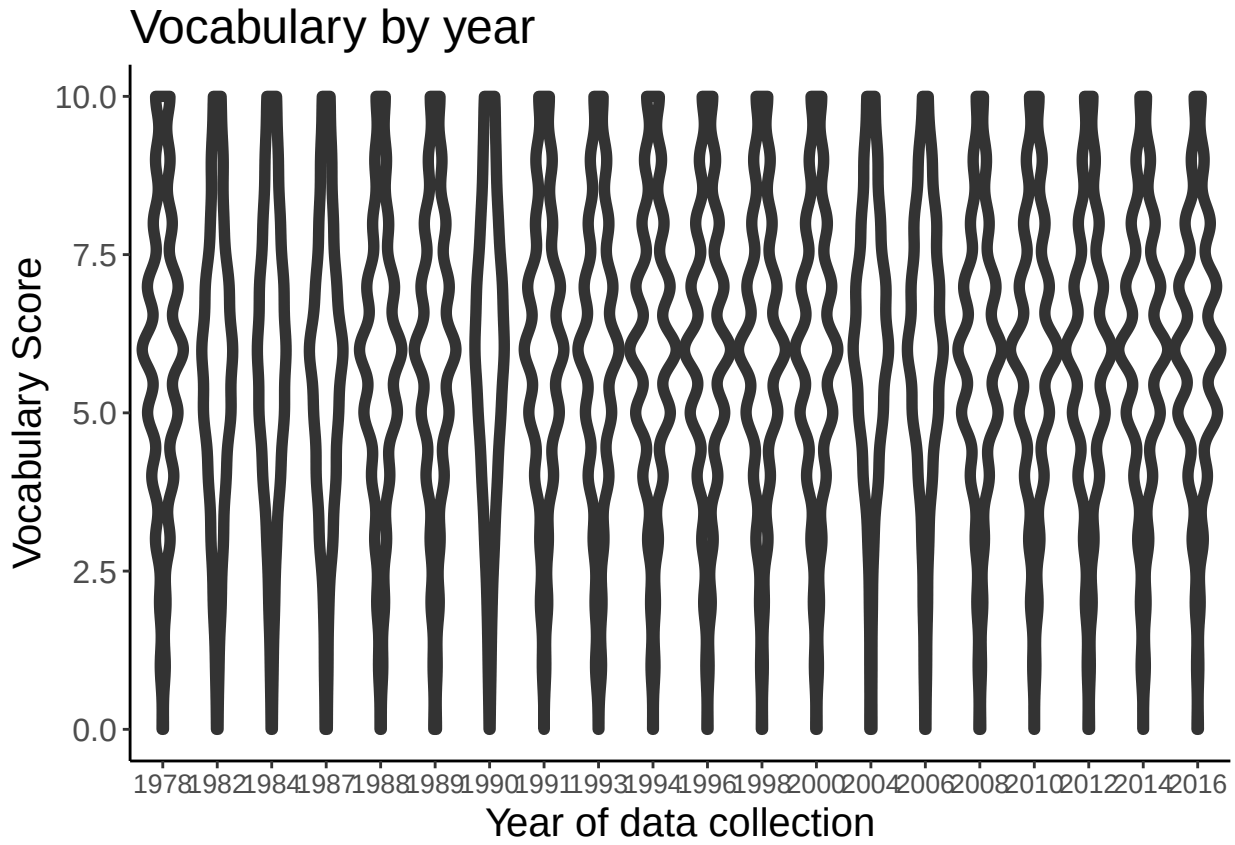
To examine our data, we can first plot it in different ways.

Let's start by looking at the vocbulary scores over the years.

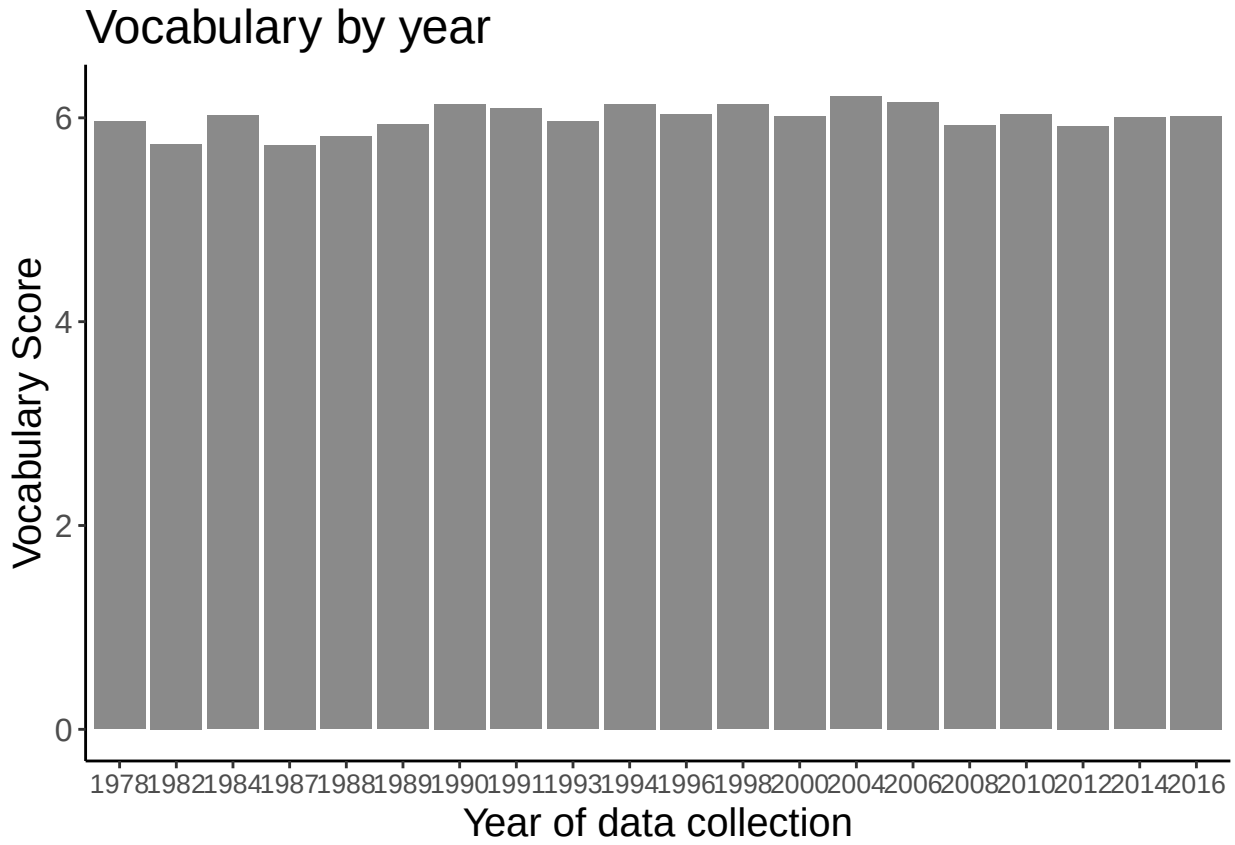
Let's try this using a box plot:



Not very useful. Maybe a violin plot?

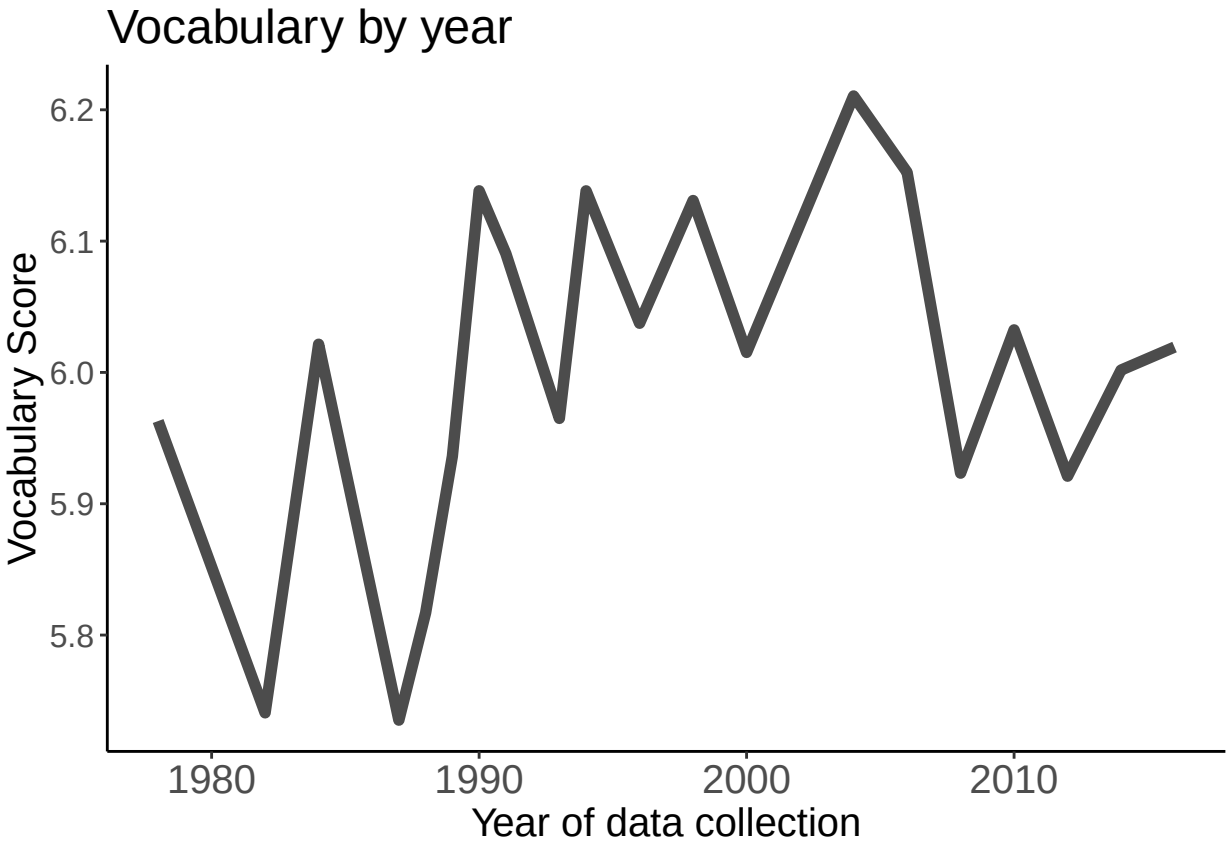


Still not great. How about a bar plot?

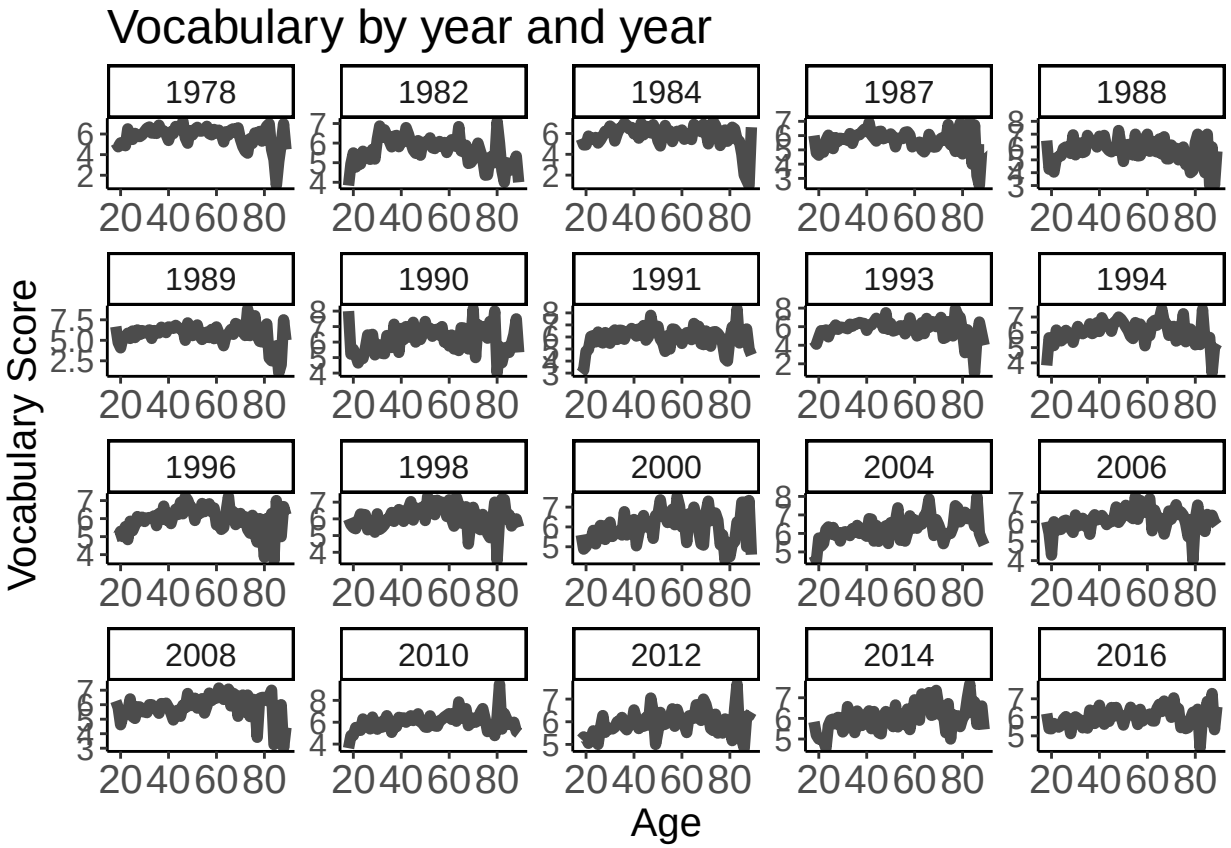


Much better, but it's still hard to read.

Perhaps it's best to summarize the scores by year, and then plot the average - that's more clean and informative.

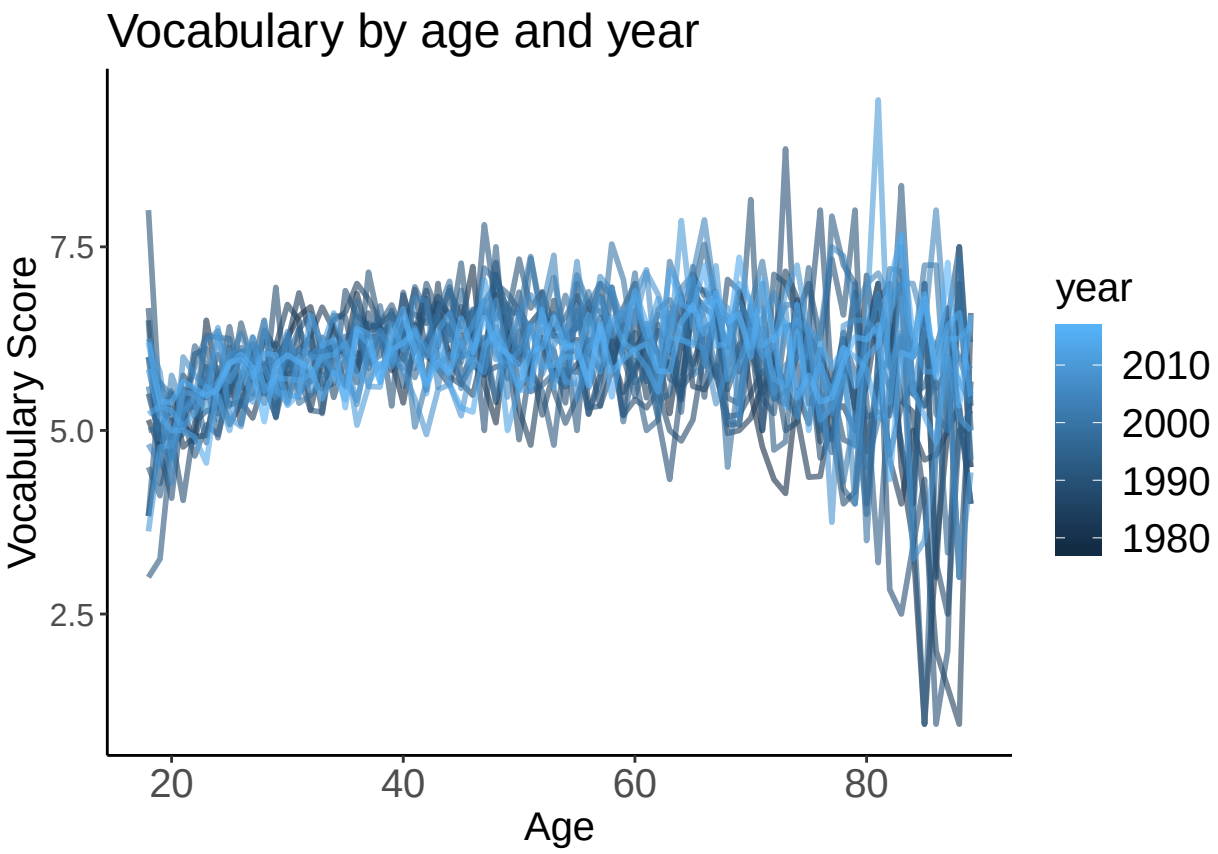


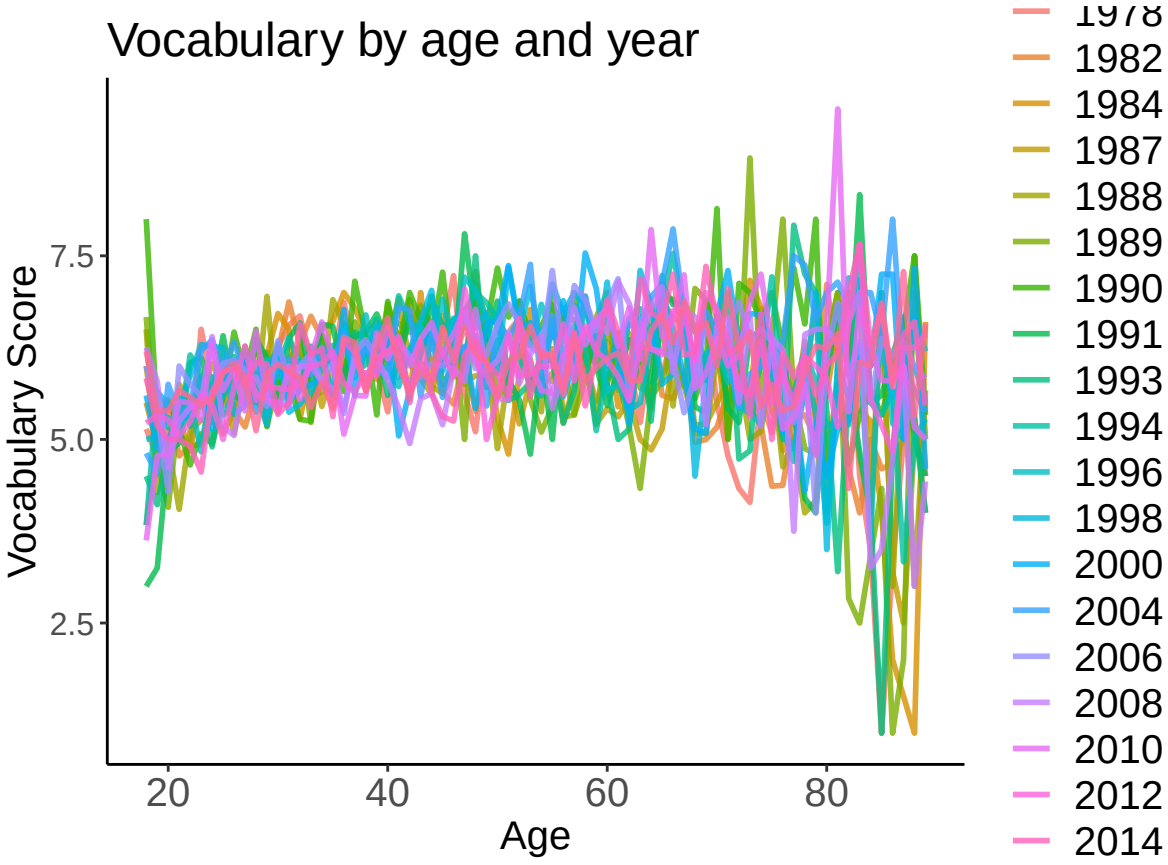
We can even examine the relationship between age and vocabulary scores over the years to make sure it's consistent. We can try to do this with faceting by year:



But this is too hard to evaluate properly. It would be better to stack the lines on top of each other and color-code them by year, so we can actually see if there are any meaningful differences. We do this by adding "color=year" to the aes().

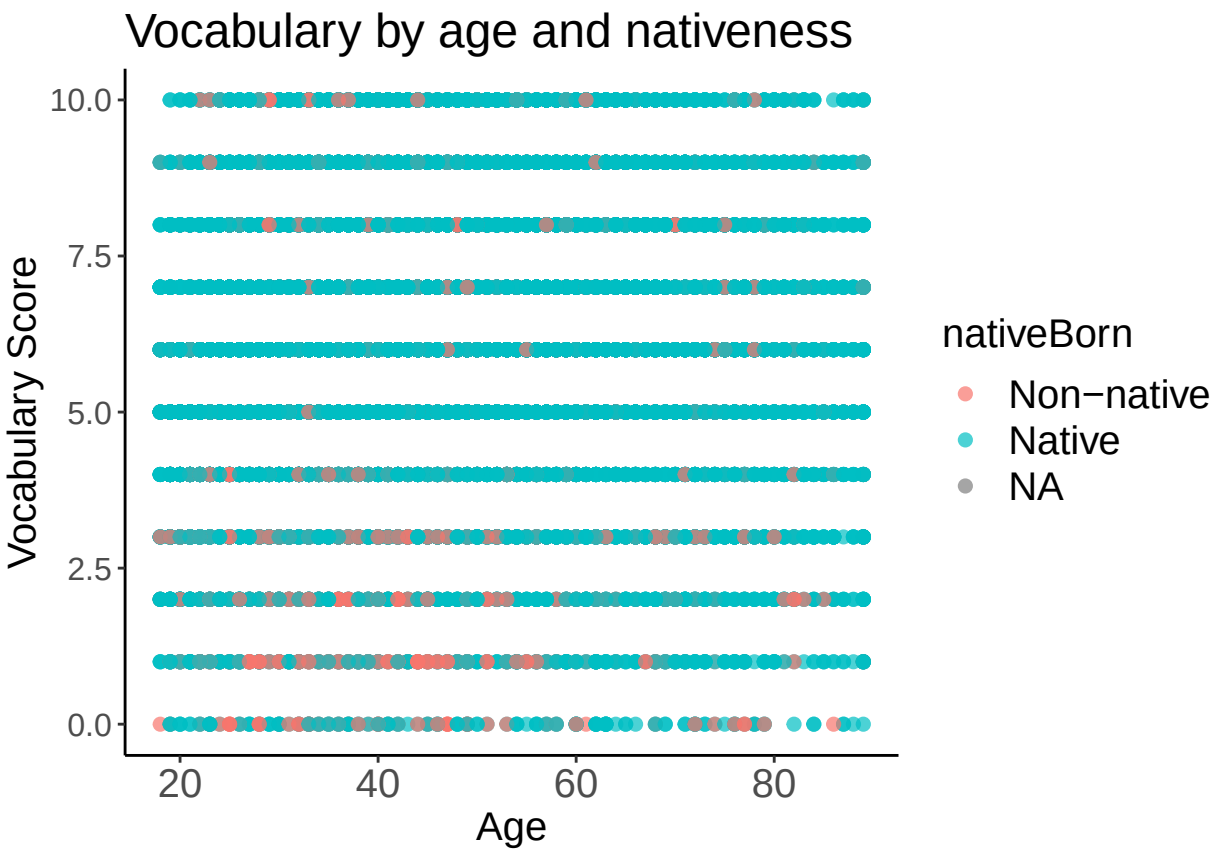
Here, we also add the "group" variable to tell ggplot to give one line per year. You can check to see what happens if you remove this grouping!


Vocabulary by age and year
Vocabulary Score
7.5
5.0
2.5
20
40
60
80
Age
year
2010
2000
1990
1980



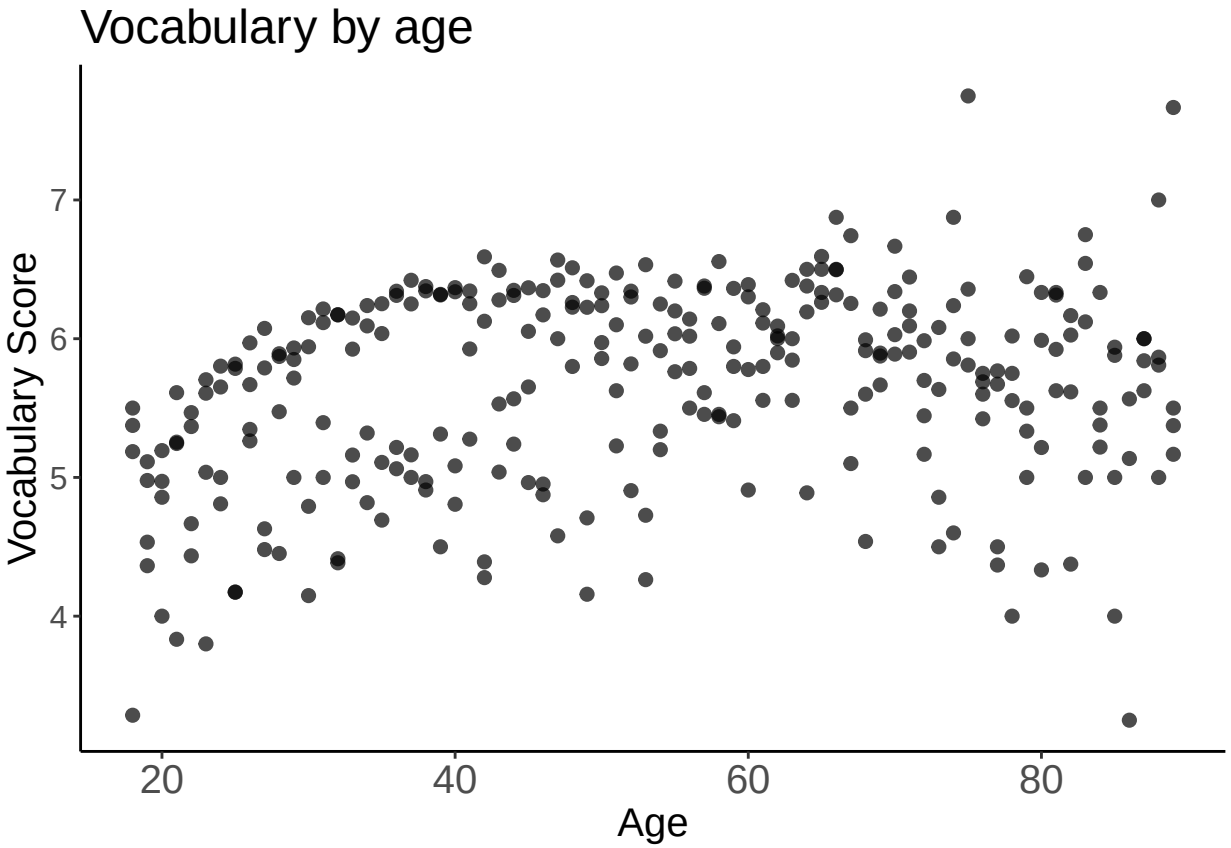
Now, let's move on and examine with the basic relation between age and vocabulary scores in our data set:

```
## Warning: Removed 65 rows containing missing values (geom_point).
```

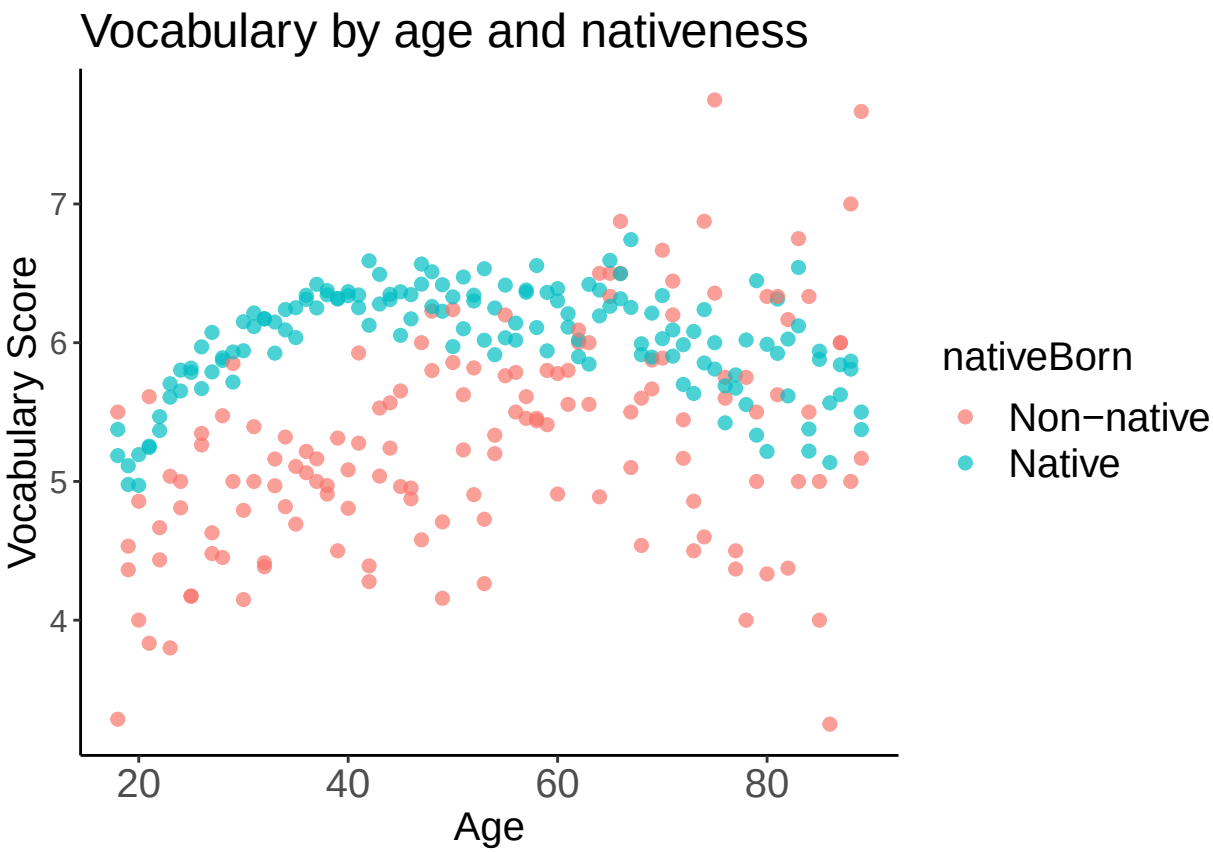


OK, this is obviously not very useful. There are just too many points!

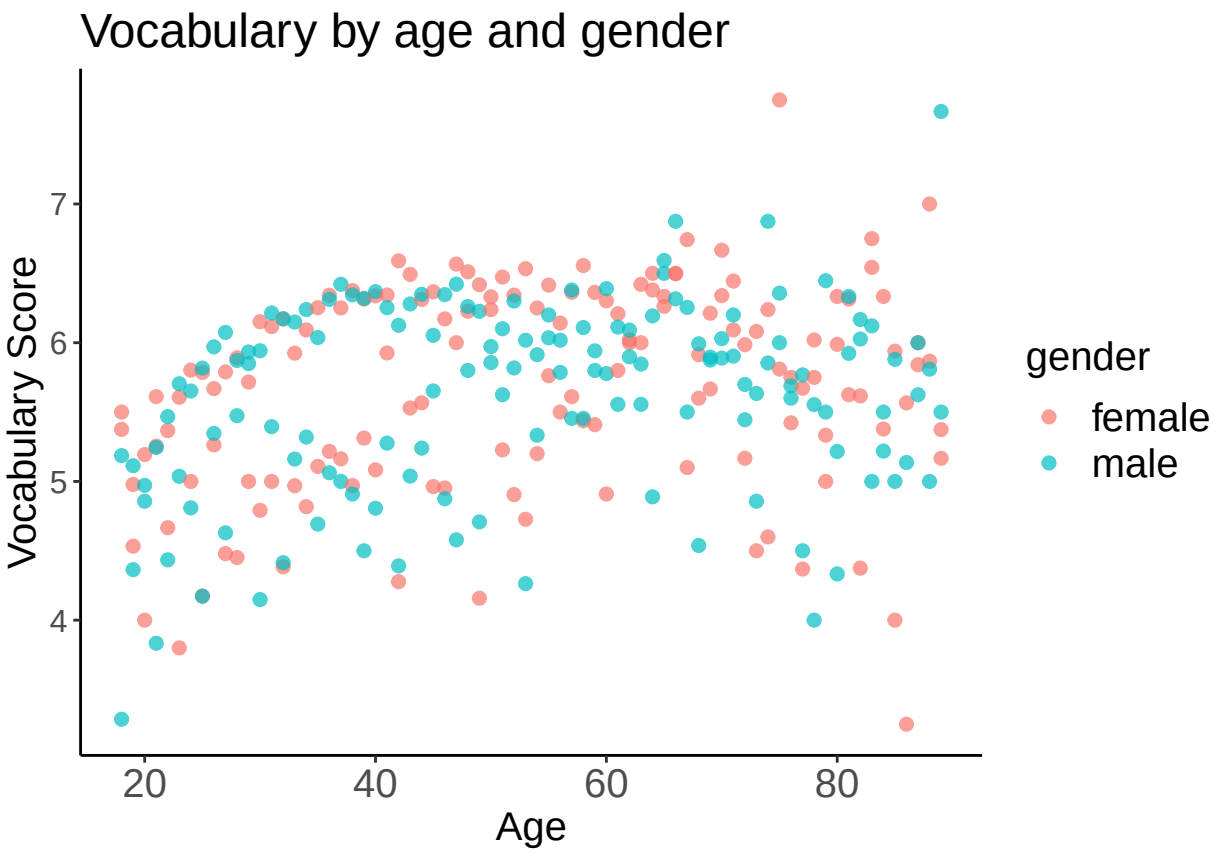
So, we can summarize the data and look at the averages.



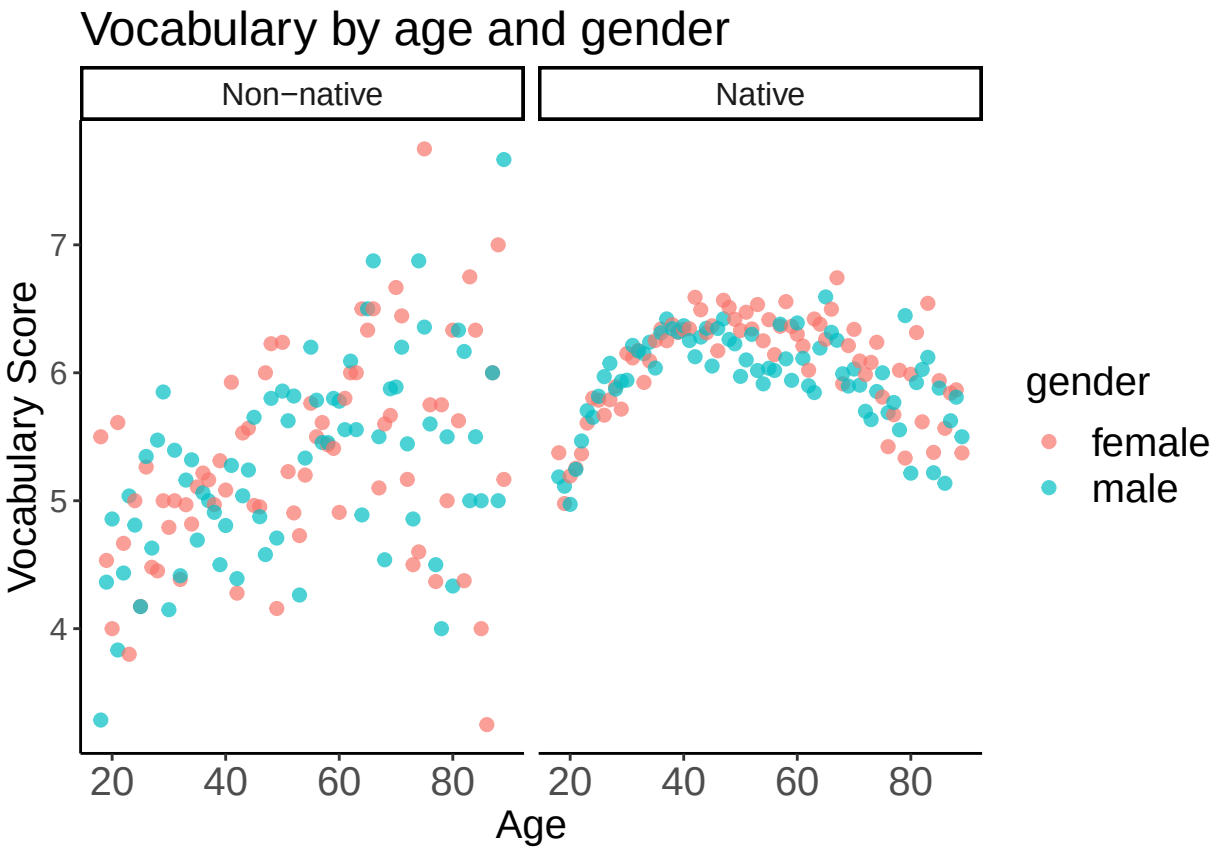
But what about being a native? Could this moderate the effect? Let's check this by adding a different color to the plot based on nativeness.

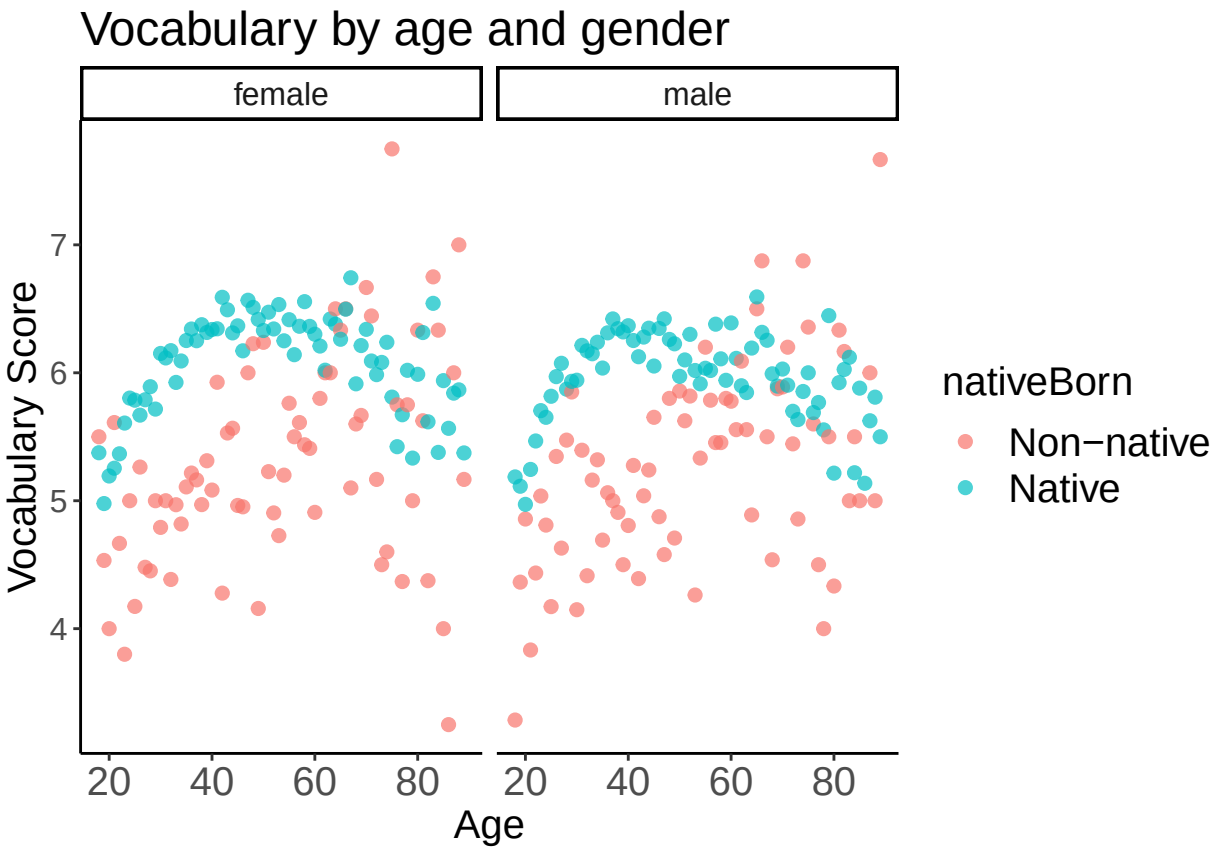


What about gender? Does it affect the relationship between age and vocabulary size?



Now, we can make more informative plots by faceting:


Vocabulary by age and gender
Non−native
Native
Vocabulary Score
7
6
5
4
20
40
60
80
Age
gender
female
male



What do you see? Do nativeness and gender have similar effects?

Time to do some plotting yourself! :)

1. Creat a new r code chunck called plot5
2. In it, make a new file that summarizes the data frame by education, gender and nativeness
3. Plot vocabulary scores by education and nativeness (same as plot2, but don't forget to change the name of the axis and the title!)
4. Plot vocabulary scores by education and gender (same as plot2a, but don't forget to change the name of the axis and the title!)
5. Facet these plots by either nativeness or gender (same as plot3 and plot3a, but don't forget to change the name of the axis and the title!)
6. Choose one of your plots and change the size of the dots, their transperancy and the size of the text.
7. Choose one of your plots and change the shape of the point by adding "shape=XX". See the legend below for help.
8. Choose the plot you made in (3), and try to make the shape of the point change according to nativness.

You can also combine different plots to one grid using the "cowplot" package. For that, you'll need the plots you saved:

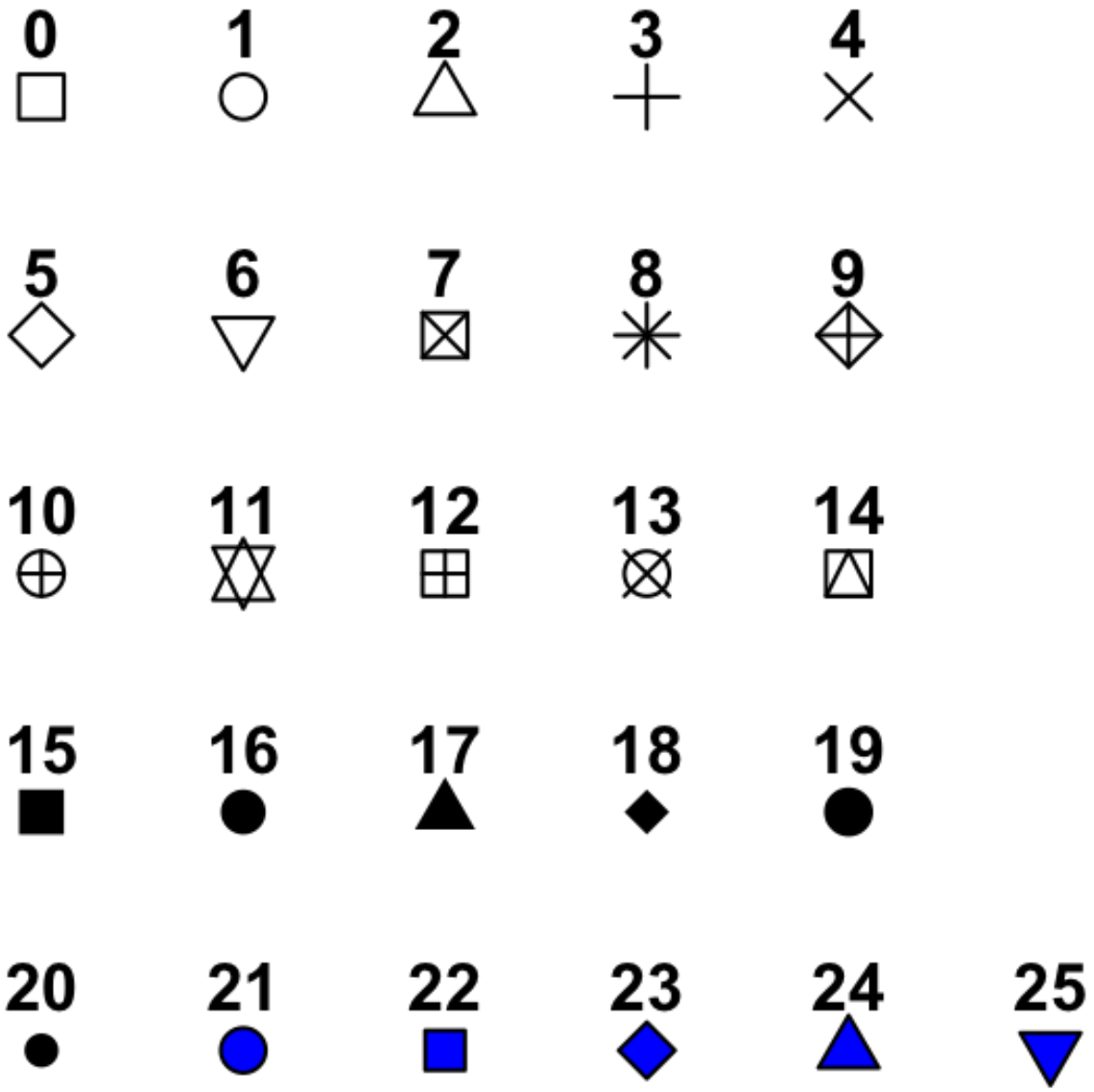
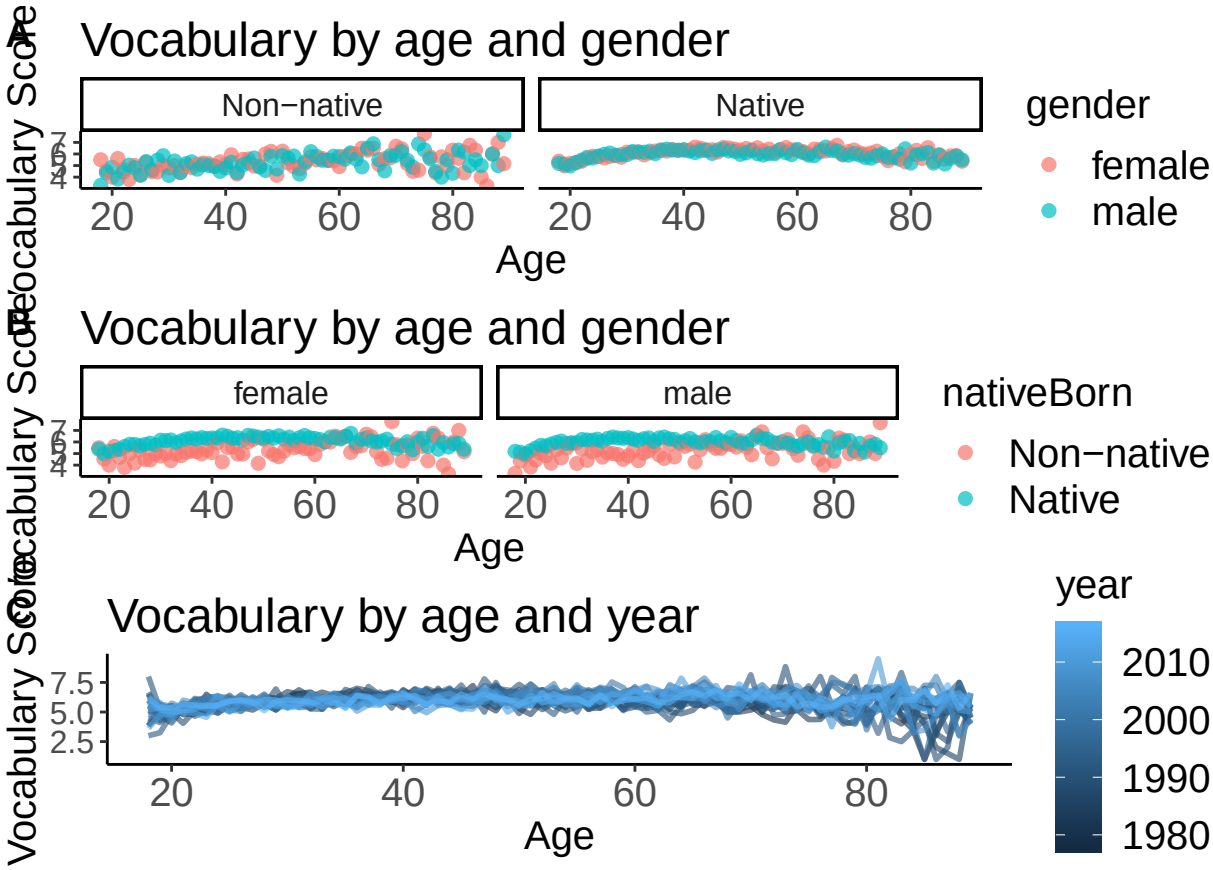
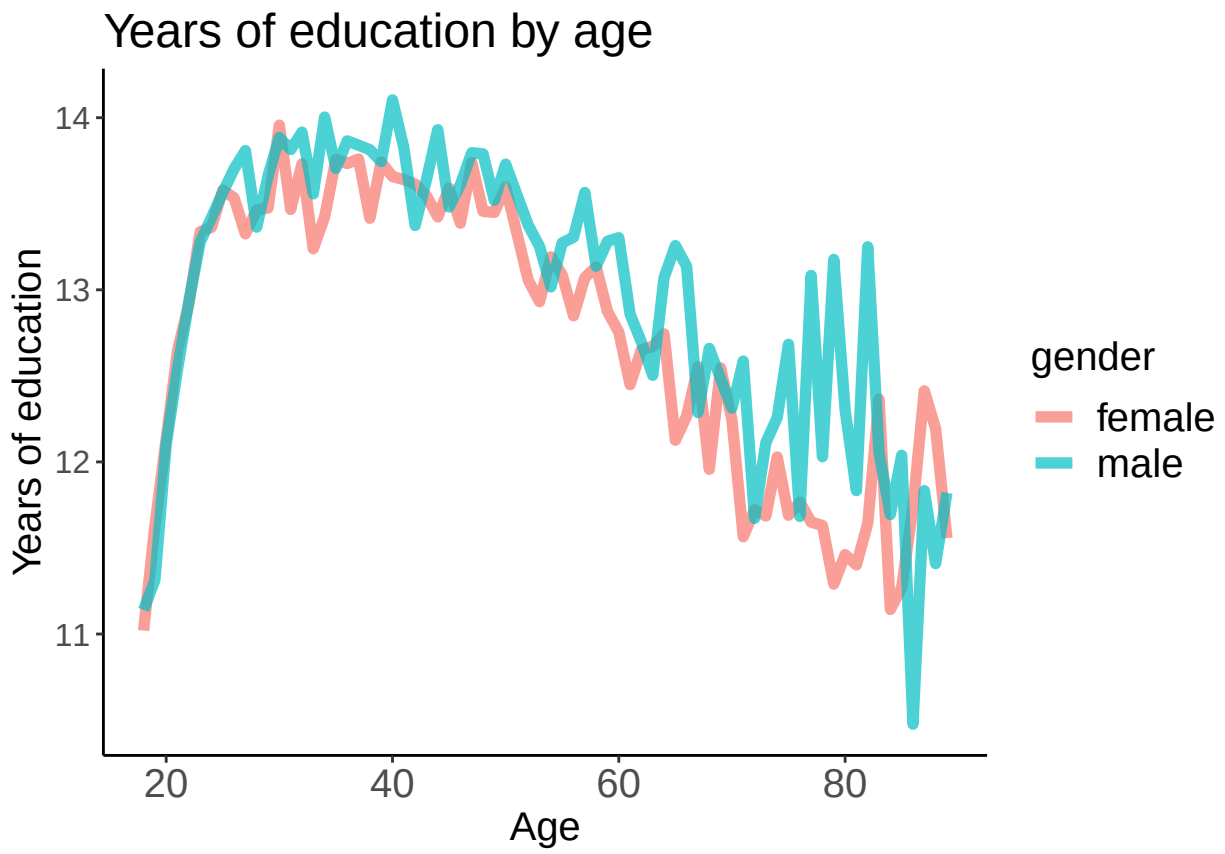


Figure 1:



We can also check our idea about a correlation between the years of education and people's age.



You'll see that actually, the relationship is more complex than we thought: even though very young people (<20) indeed have less education, overall it seems like older people are less educated. This might have to do with having fewer opportunities for higher education in the past. Moreover, this gap is even bigger for women...

Analysis time!

First, let's prepare our variables by centering the continous ones and setting up the contrasts.

Now, let's make a model based on our predictions.

Note: We're not going to do model selection this time (mostly for the sake of time, and because it can take a while to find the model that actually converges and is justified by the data), but you're welcome to try other models yourself at home!

Adding p-values

Because this model is based on a lot of data, we can use the normal approximation to caluclate p-values: Since the t-distribution converges to the z-distribution as degrees of freedom increase, this is like assuming infinite degrees of freedom. For reasonable sample sizes, this method appears not to be very anti-conservative (see Barr et al., 2013: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3881361/). That is, if we take the p-value to measure the probability of a false positive, this approximation produces a somewhat (but perhaps not alarmingly) higher false positive rate than the nominal 5% at $p = 0.05$. For our "big" dataset, this is acceptable.

(I also added a code at the end of this file with a way to caluclate p-values for small sample sizes.)

Table 1: Vocabulary score by age, education, nativeness and gender

	Estimate	Std.Error	t-value	p-value
(Intercept)	6.131574	0.037432	163.803703	0.000000
Age	0.016822	0.000863	19.490161	0.000000
Nativeness (Non-native vs. Native)	-0.809289	0.053148	-15.226985	0.000000
Gender (Male vs. Female)	-0.142522	0.023059	-6.180831	0.000000
Years of Education	0.370656	0.005566	66.589170	0.000000
Age X Nativeness	0.006498	0.003223	2.016057	0.043794
Age X Gender	-0.006632	0.001341	-4.946599	0.000001
Nativeness X Gender	0.066434	0.079227	0.838532	0.401732
Nativeness X Education	-0.113665	0.014206	-8.001045	0.000000
Education X Gender	0.000753	0.007976	0.094451	0.924751
Age X Nativeness X Gender	0.010527	0.004842	2.174262	0.029686
Education X Nativeness X Gender	-0.025106	0.020406	-1.230326	0.218575

But what's actually going on?

It's fairly easy to understand the main effects of education, age, gender and nativeness:

- Age is a significant positive predictor of vocab scores (higher age = higher score)
- Nativeness is a significant negative predictor of vocab scores (non native = lower score than native)
- Gender is a significant negative predictor of vocab scores (males = lower score than females)
- Education is a significant positive predictor of vocab scores (higher education = higher score)

But when it comes to the interactions (and especially, the triple interaction), things get messier.

We can try to use the sign of the interactions to understand our effects.

For exmaple: - the effect of age is positive (higher age = higher score) - the effect of gender is negative (males = lower score) - the interaction between age and gender is negative (–> the positive effect of age on vocab score is smaller for males)

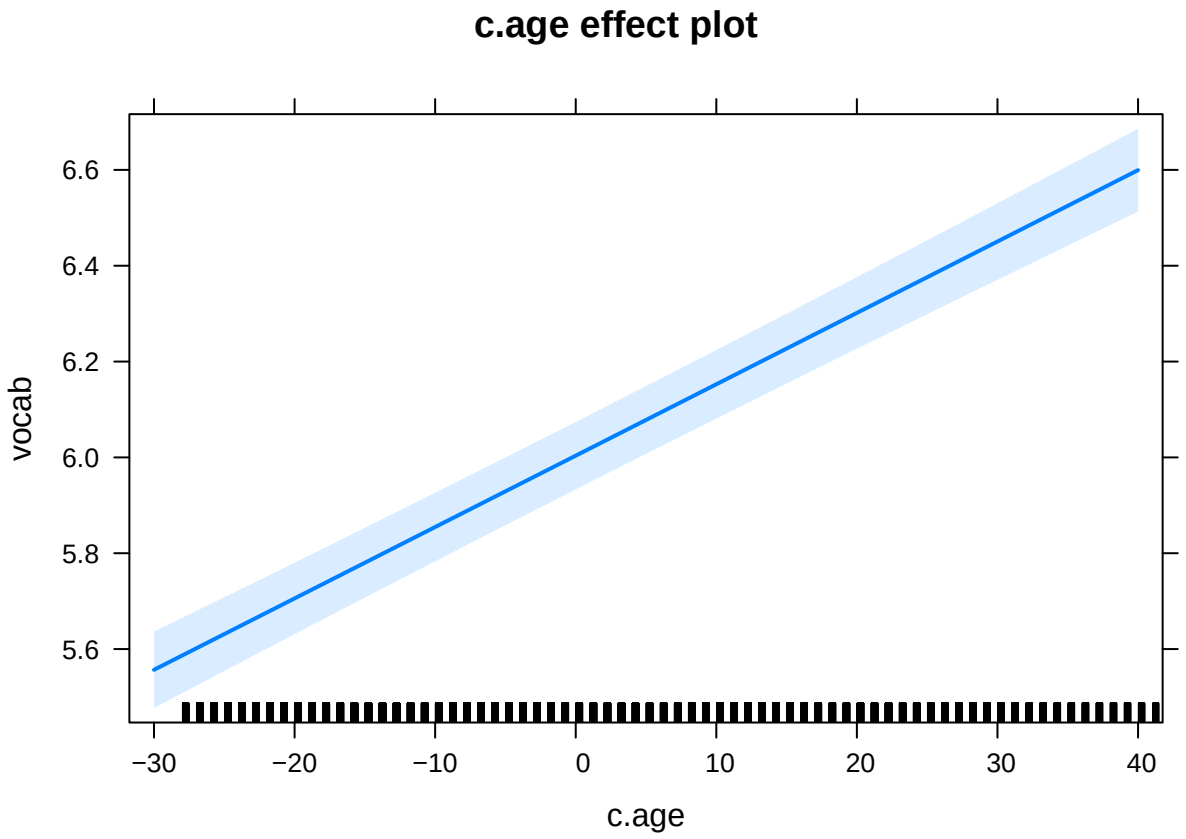
But this is not trivial for everyone, and can be confusing.

So - we plot the model using the "effects" package! :)

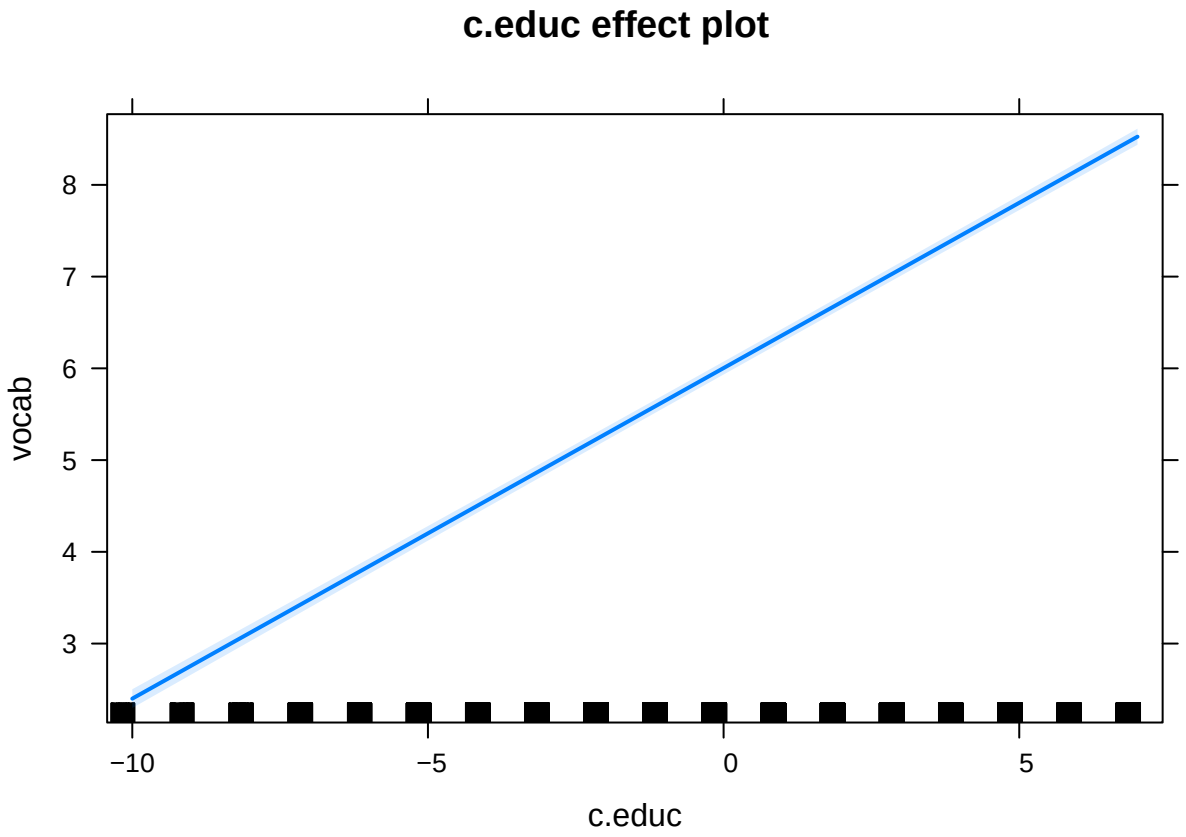
Plotting the model

First, let's confirm that we understood the main effects:

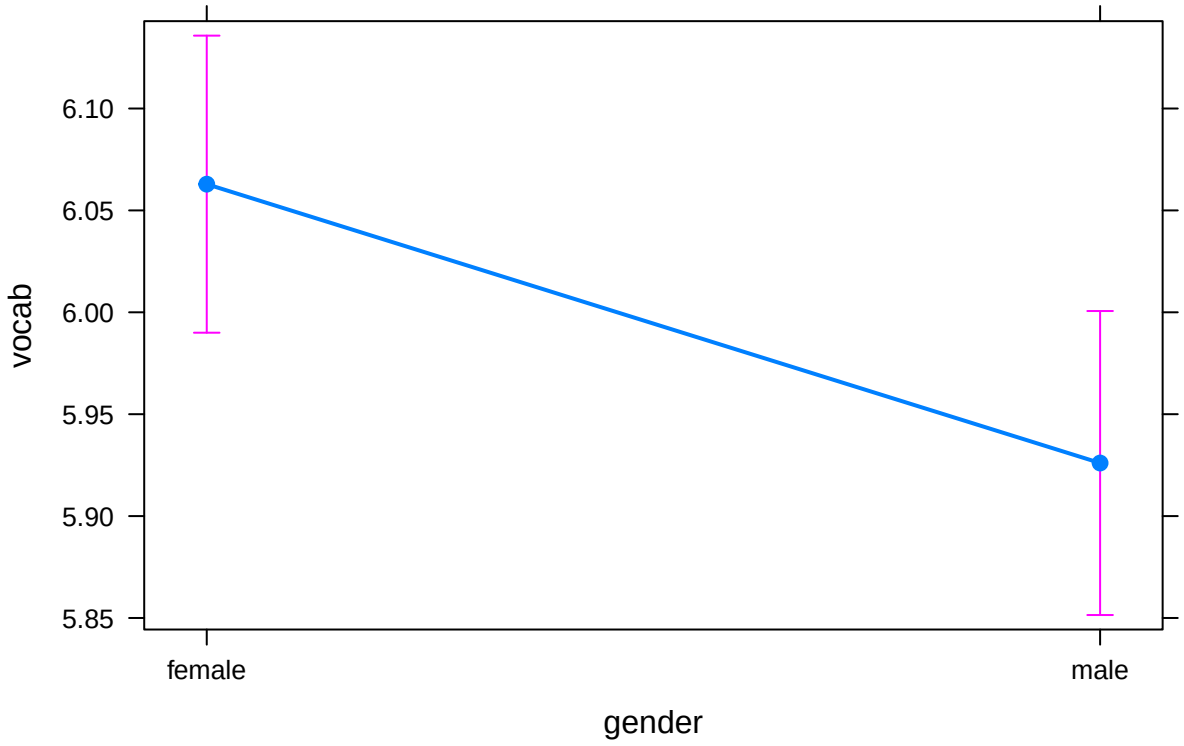
```
## NOTE: c.age is not a high-order term in the model
```



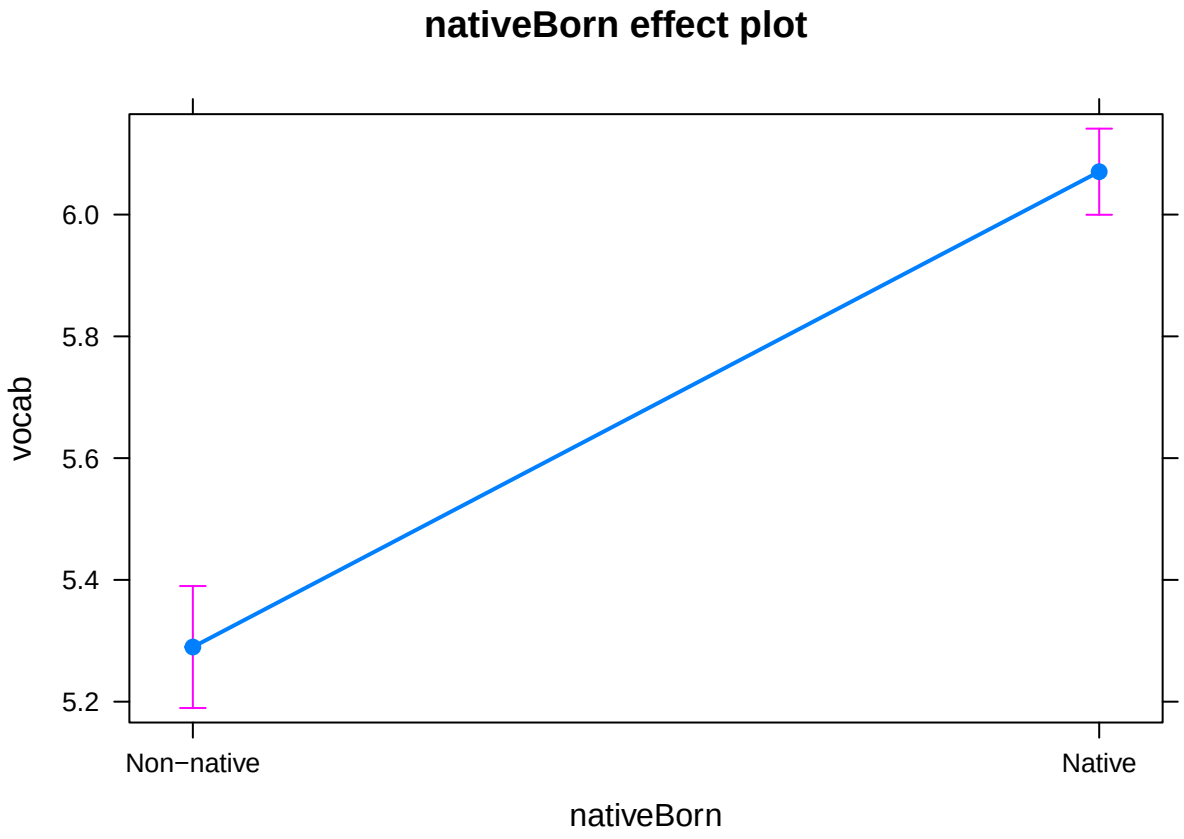

```
## NOTE: c.educ is not a high-order term in the model
```




```
## NOTE: gender is not a high-order term in the model
```



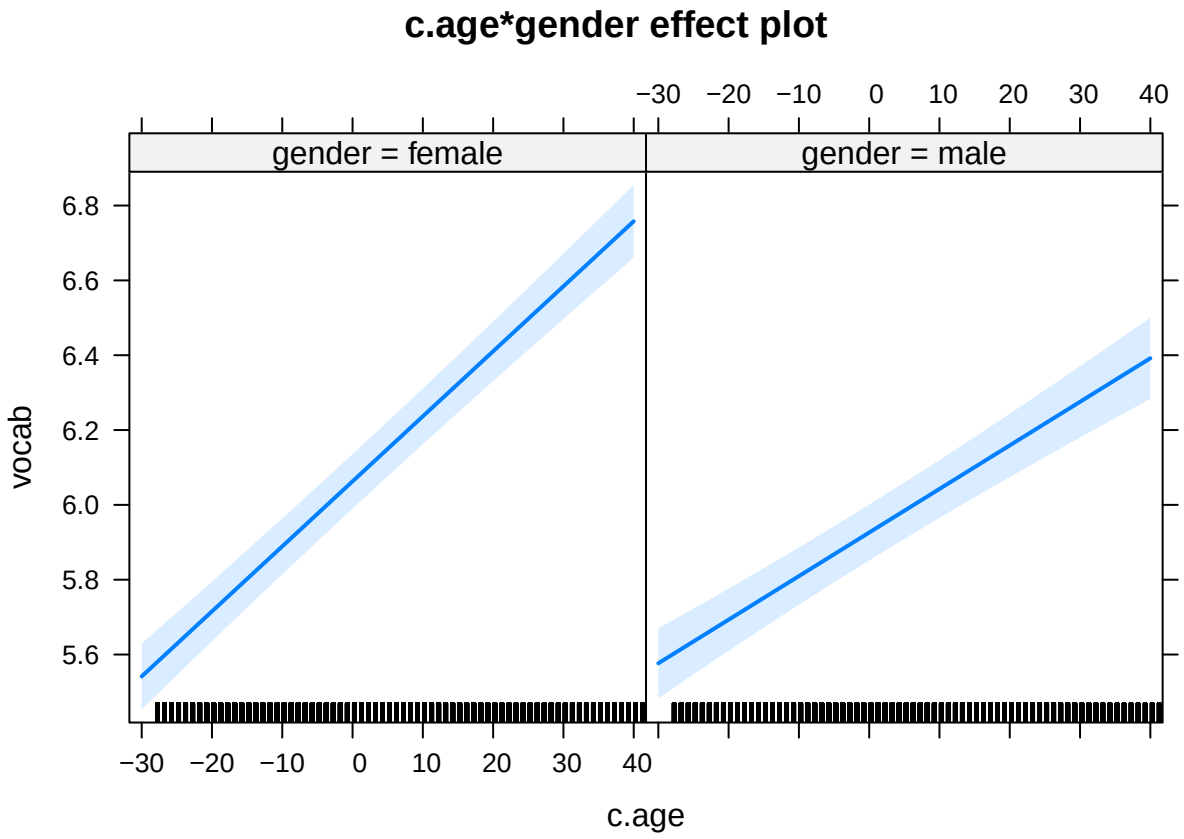

```
## NOTE: nativeBorn is not a high-order term in the model
```



Now let's plot the interactions:

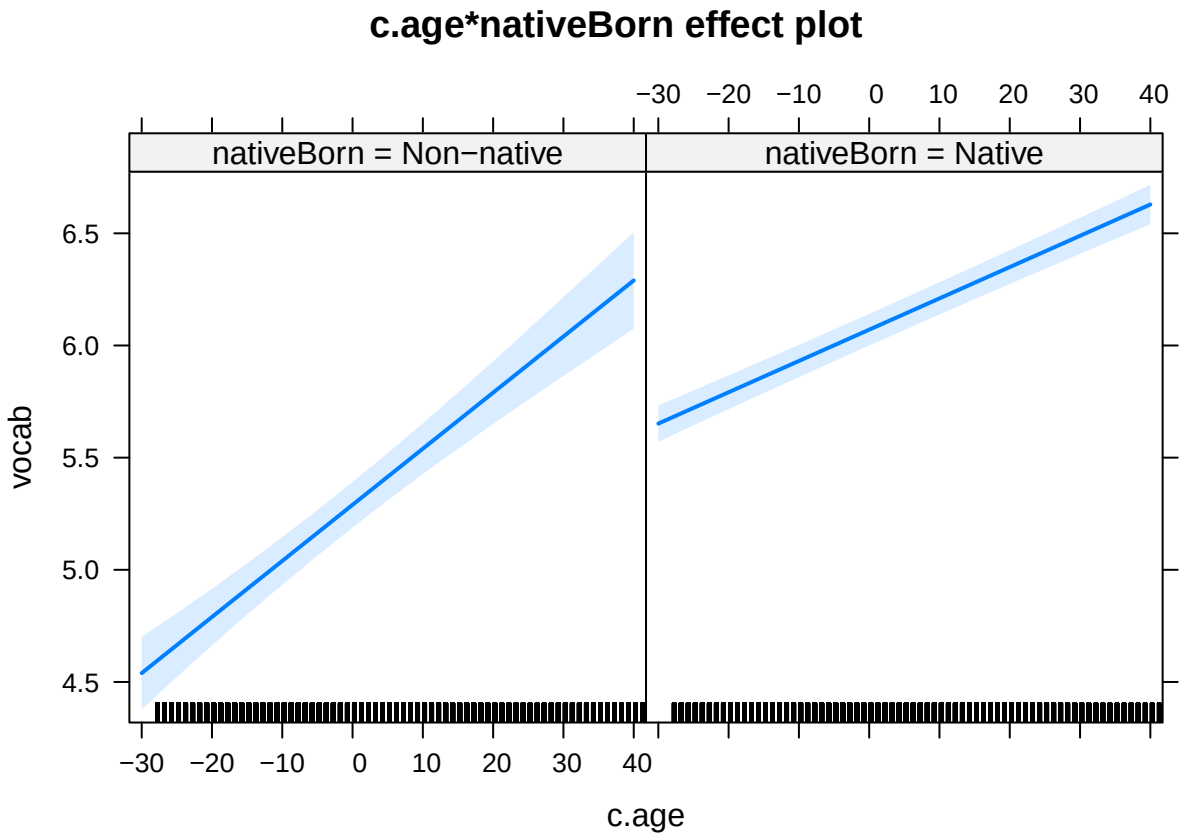
Note: The name of the effect needs to be according to the order in the original model)

```
## NOTE: c.age:gender is not a high-order term in the model
```



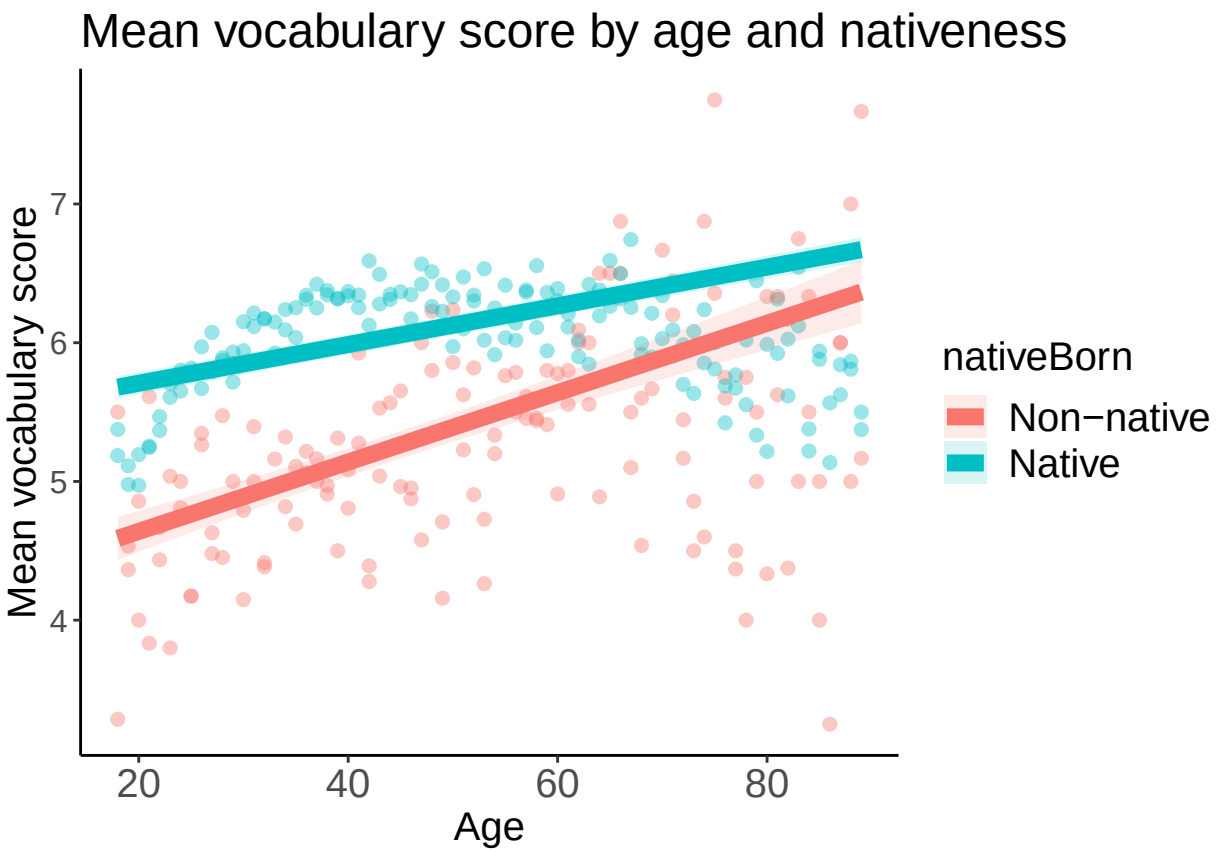
Indeed, we can see that the relationship between age and vocabulary scores is weaker (=smaller slope) for males.

```
## NOTE: c.age:nativeBorn is not a high-order term in the model
```



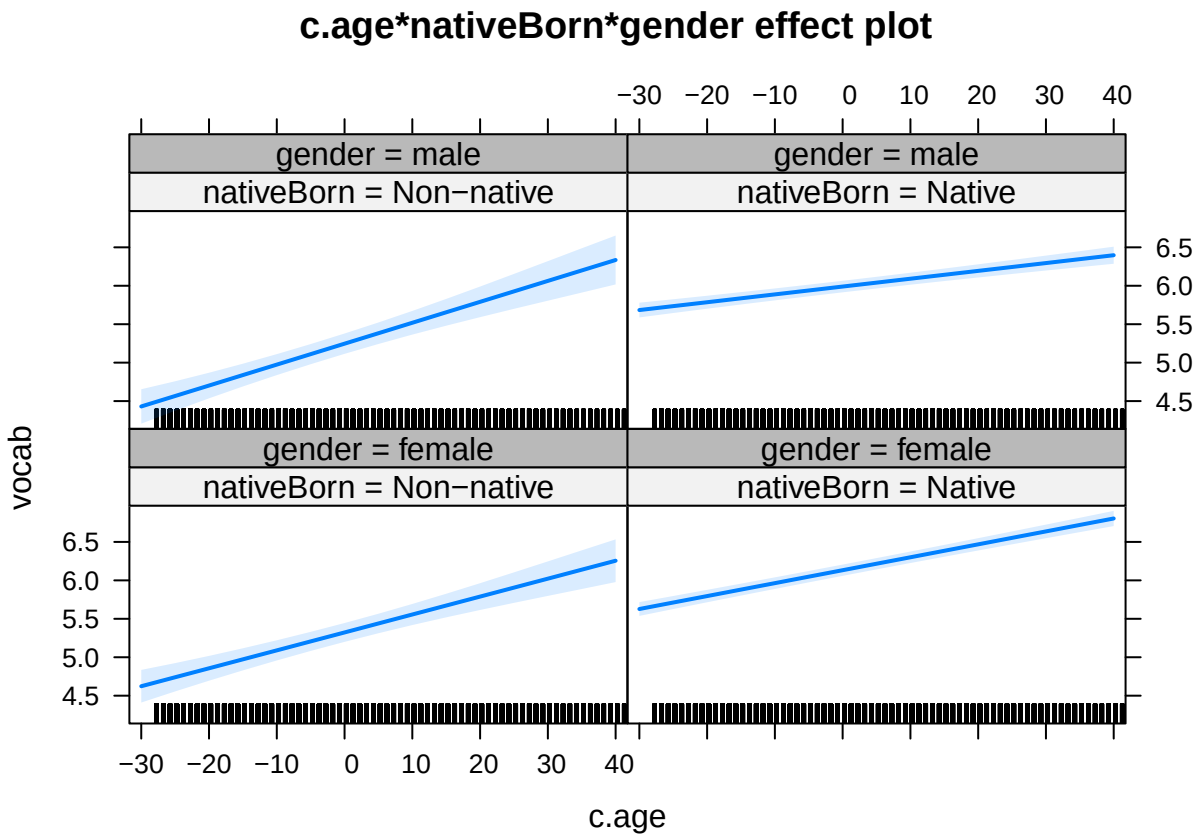
It might be useful to plot these slopes from the model on top on the raw data. This is actually easy to do in ggplot: we can combine data from different files together.

```
## NOTE: c.age:nativeBorn is not a high-order term in the model
```



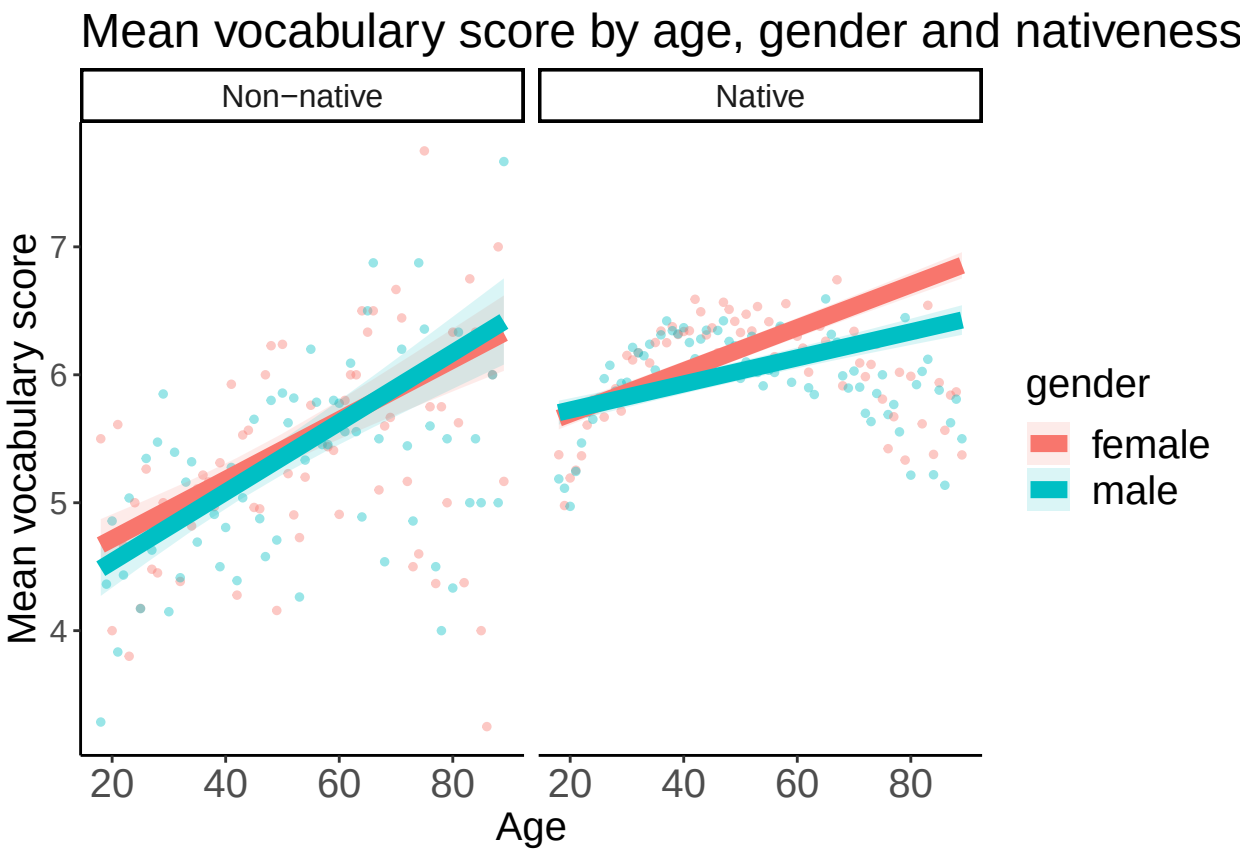
Great! We see that except for the fact that natives have higher scores, nativeness is changing the relationship between age and vocabulary scores (i.e., the slope is steeper for non-natives).

Plotting (and understanding) triple interactions



This plot is actually not so clear and it's hard to see the differences, but the idea is that the significant relationship between age and gender (i.e., males showing a weaker relationship) is somewhat modluated by nativeness (so it is true only for the natives).

Maybe we can try to plot the model AND the raw data together to get a better picture.



Time to do some model plotting yourself! :)

1. Creat a new r code chunck called regression plot7
2. Plot the interaction between nativeness and gender
3. Plot the interaction between nativeness and education
4. Plot the interaction between gender and education
5. Plot the triple interaction between gender, education and nativeness
6. Plot the raw data + model estimates for one of the interactions above
7. Check: do the plots fit the model's output and the significance of the interactions?

Side note about contrasts for categorical variables:

The "basic" coding schemes are: - *Dummy coding* = compares levels to a baseline level (this is done with 0's and 1's) - *Sum/deviation conding* = compares levels to the grand mean (this is done with -1's and 1's)

There are *many* other coding schemes to make many more comparisons (even user-defined ones). You can check out this webpage for useful exmaples and coding matrices: https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/

Using different conding schemes asks different questions, and makes different comparisons. It will therefore yield different coefficients (estimates). *But the overall model fit stays the same, and* ***the significance of the predictor remains the same!!***

If you want to be sure, let's rerun the same model after changing the contrasts of "nativeBorn" to c(-1,1) and check for yourself!

Table 2: Original model

	Estimate	Std.Error	t-value	p-value
(Intercept)	6.131574	0.037432	163.803703	0.000000
Age	0.016822	0.000863	19.490161	0.000000
Nativeness (Non-native vs. Native)	-0.809289	0.053148	-15.226985	0.000000
Gender (Male vs. Female)	-0.142522	0.023059	-6.180831	0.000000
Years of Education	0.370656	0.005566	66.589170	0.000000
Age X Nativeness	0.006498	0.003223	2.016057	0.043794
Age X Gender	-0.006632	0.001341	-4.946599	0.000001
Nativeness X Gender	0.066434	0.079227	0.838532	0.401732
Nativeness X Education	-0.113665	0.014206	-8.001045	0.000000
Education X Gender	0.000753	0.007976	0.094451	0.924751
Age X Nativeness X Gender	0.010527	0.004842	2.174262	0.029686
Education X Nativeness X Gender	-0.025106	0.020406	-1.230326	0.218575

Table 3: Model with sum-coding

	Estimate	Std.Error	t-value	p-value
(Intercept)	5.726930	0.043339	132.144147	0.000000
Age	0.020071	0.001612	12.450813	0.000000
Nativeness (Non-native vs. Native)	-0.404644	0.026574	-15.226986	0.000000
Gender (Male vs. Female)	-0.109305	0.039611	-2.759472	0.005789
Years of Education	0.313824	0.007114	44.115960	0.000000
Age X Nativeness	0.003249	0.001612	2.016057	0.043794
Age X Gender	-0.001369	0.002421	-0.565363	0.571827
Nativeness X Gender	0.033217	0.039613	0.838532	0.401732
Nativeness X Education	-0.056832	0.007103	-8.001044	0.000000
Education X Gender	-0.011800	0.010206	-1.156118	0.247633
Age X Nativeness X Gender	0.005264	0.002421	2.174262	0.029686
Education X Nativeness X Gender	-0.012553	0.010203	-1.230326	0.218575

Add conservative p-values to small models

Finally, if you have a small data set (e.g., less than 30 participants, or very few observation per condition), you can use the *Kenward-Roger approximation* for estimating the degrees of freedom in a mixed model. This method is slightly more conservative.

Let's try this using a model based on a smaller sample size (for example, with only part of the data from the last two years of data collection).

Note: If your model is not *that* small, this procedure might take a while to calculate and can be quite taxing for your computer.

Table 4: Comparing p-values for small model

	Estimate	Std.Error	t-value	p-value (KR)	p-value (normal)
(Intercept)	5.909471	0.309299	19.106033	0.000001	0.000000

	Estimate	Std.Error	t-value	p-value (KR)	p-value (normal)
Age	0.028100	0.019589	1.434484	0.200601	0.151434
Nativeness (Non-native vs. Native)	-0.412873	0.309299	-1.334867	0.229533	0.181920
Age X Nativeness	0.014420	0.019589	0.736143	0.488947	0.461644

As you can see, using the KR is a bit more conservative (the p-values are slightly higher).