Assignment2_CDS

sofiascharf

2026-01-07

Part 1

Load dataset

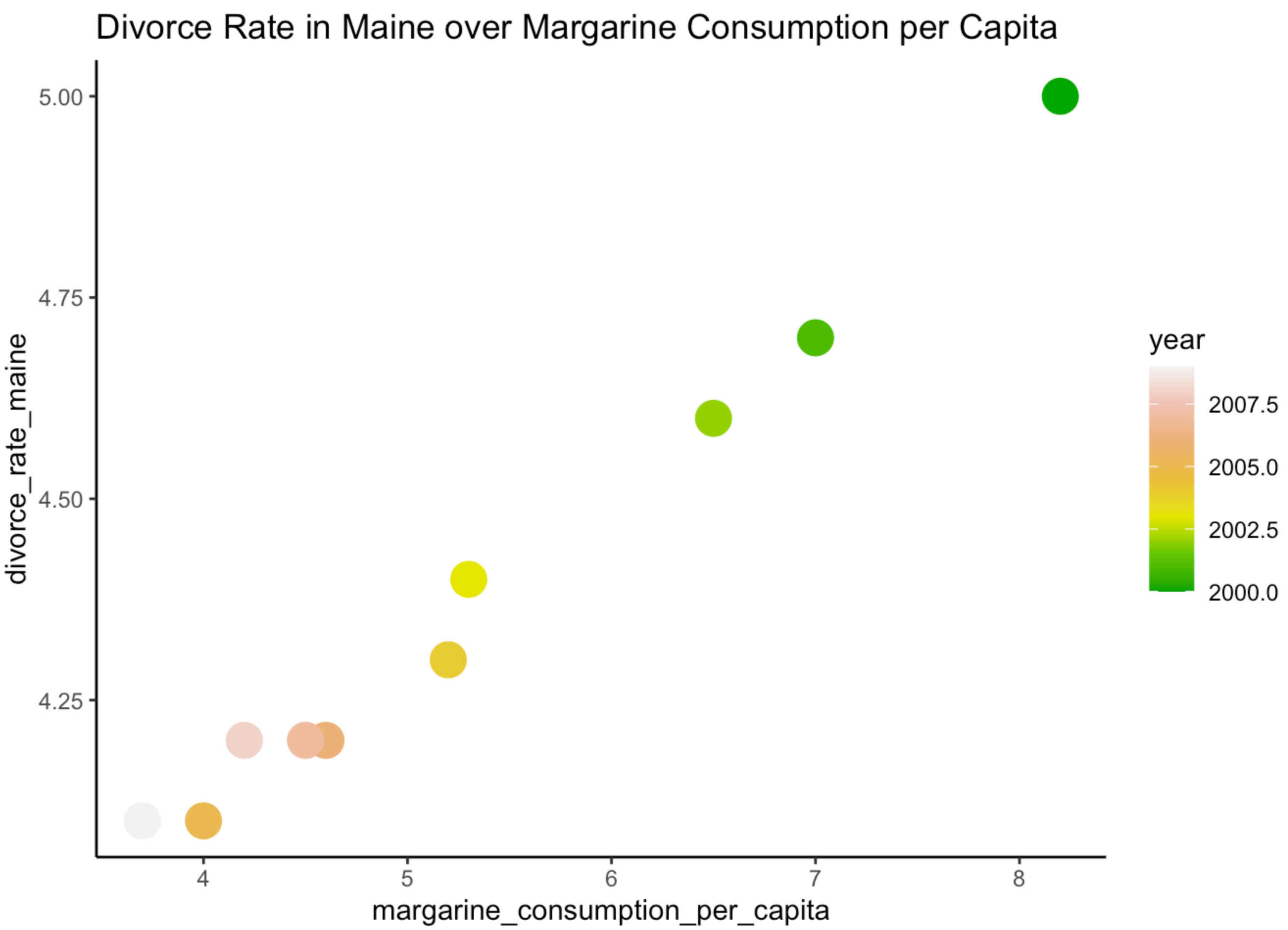
```
divorce <- dslabs::divorce_margarine
head(divorce)
```

```
##   divorce_rate_maine margarine_consumption_per_capita year
## 1                5.0                              8.2 2000
## 2                4.7                              7.0 2001
## 3                4.6                              6.5 2002
## 4                4.4                              5.3 2003
## 5                4.3                              5.2 2004
## 6                4.1                              4.0 2005
```

Plot

```
ggplot(divorce, # init plot and choose axes n df
       aes(x = margarine_consumption_per_capita,
           y = divorce_rate_maine,
           color = year)) +
  geom_point(size = 6) + # point size
  scale_colour_gradientn(colours = terrain.colors(7)) + # set palette
  ggtitle("Divorce Rate in Maine over Margarine Consumption per Capita") +
  theme_classic()
```

Loading [MathJax]/jax/output/HTML-CSS/jax.js



Already looks like there could be a linear relationship.

Stats

```
cor.test(divorce$divorce_rate_maine,
    divorce$margarine_consumption_per_capita,
    method = "spearman") # continous variables
```

```
## Warning in cor.test.default(divorce$divorce_rate_maine,
## divorce$margarine_consumption_per_capita, : Cannot compute exact p-value with
## ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  divorce$divorce_rate_maine and divorce$margarine_consumption_per_capita
## S = 2.5192, p-value = 2.334e-07
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9847319
```

According to the results of our correlations test, there is a strong, significant linear relationship between

Loading [MathJax]/jax/output/HTML-CSS/jax.js

margarine consumption per capita and divorce rate in Maine (*rho = 0.98*, *p < .001*).

Part 2

Load and subset

```
vocab <- carData::GSSvocab
head(vocab)
```

```
##        year gender nativeBorn ageGroup educGroup vocab age educ
## 1978.1 1978 female        yes    50-59    12 yrs    10  52   12
## 1978.2 1978 female        yes      60+   <12 yrs     6  74    9
## 1978.3 1978   male        yes    30-39   <12 yrs     4  35   10
## 1978.4 1978 female        yes    50-59    12 yrs     9  50   12
## 1978.5 1978 female        yes    40-49    12 yrs     6  41   12
## 1978.6 1978   male        yes    18-29    12 yrs     6  19   12
```

```
vocab <- vocab %>%
  filter(year == 1978, fill.NA=TRUE)
head(vocab)
```

```
##        year gender nativeBorn ageGroup educGroup vocab age educ
## 1978.1 1978 female        yes    50-59    12 yrs    10  52   12
## 1978.2 1978 female        yes      60+   <12 yrs     6  74    9
## 1978.3 1978   male        yes    30-39   <12 yrs     4  35   10
## 1978.4 1978 female        yes    50-59    12 yrs     9  50   12
## 1978.5 1978 female        yes    40-49    12 yrs     6  41   12
## 1978.6 1978   male        yes    18-29    12 yrs     6  19   12
```

Initial plot

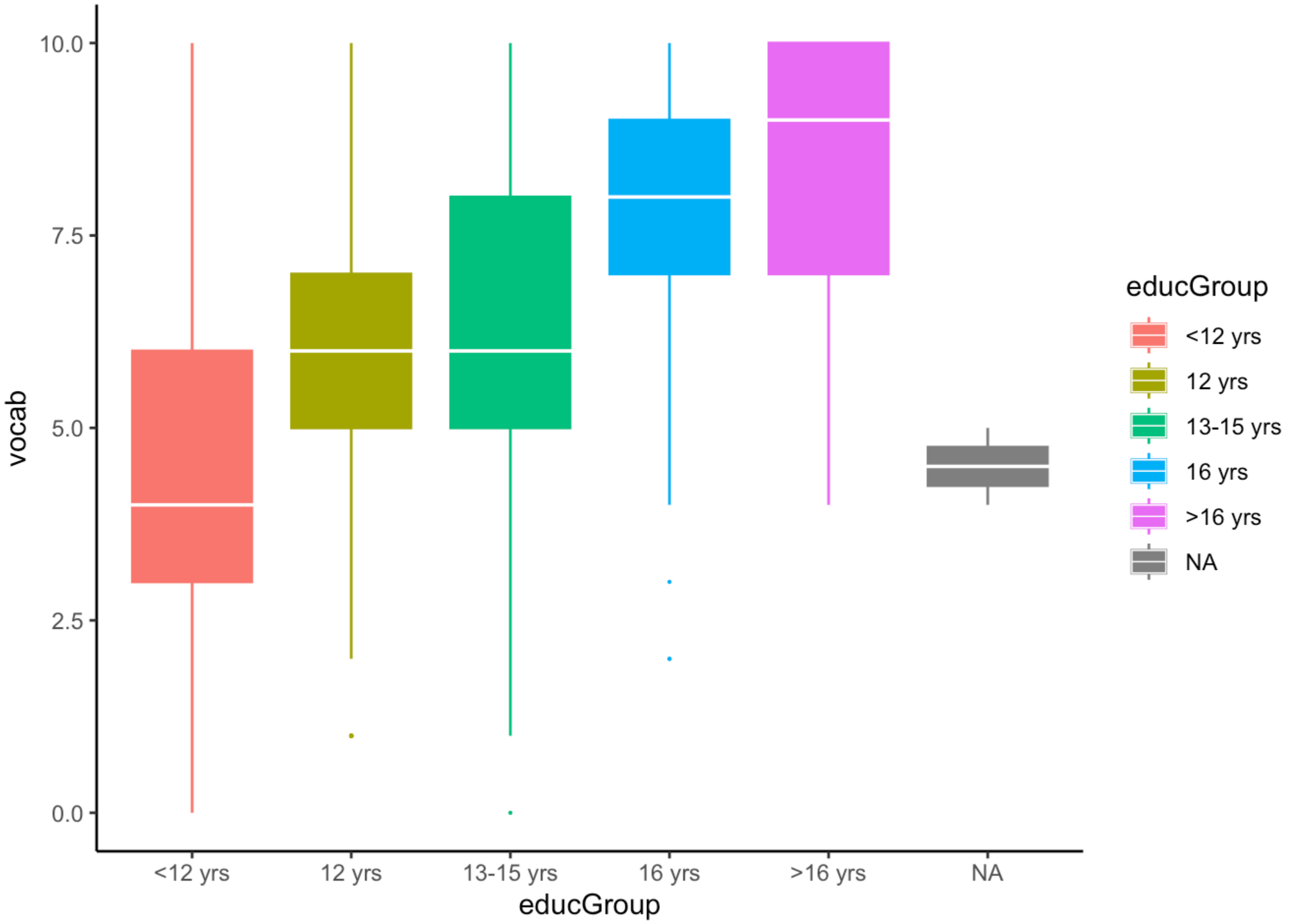
```
ggplot(vocab, # init plot and choose axes n df
       aes(x = educGroup,
           y = vocab,
           col = educGroup,
           fill = educGroup)) +
  geom_boxplot(outlier.size = 0.1) +
  theme_classic() +
  stat_summary(
    fun = median,
    geom = "crossbar",
    aes(group = educGroup),
    color = "white",       #  color for the median line
    size = 0.25,            # Set the thickness of the median line.
    width = 1          # Set the width of the median line
  )
```

Loading [MathJax]/jax/output/HTML-CSS/jax.js

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: Removed 46 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

```
## Warning: Removed 46 rows containing non-finite outside the scale range
## (`stat_summary()`).
```



Here, it looks like there could be a linear relationship.

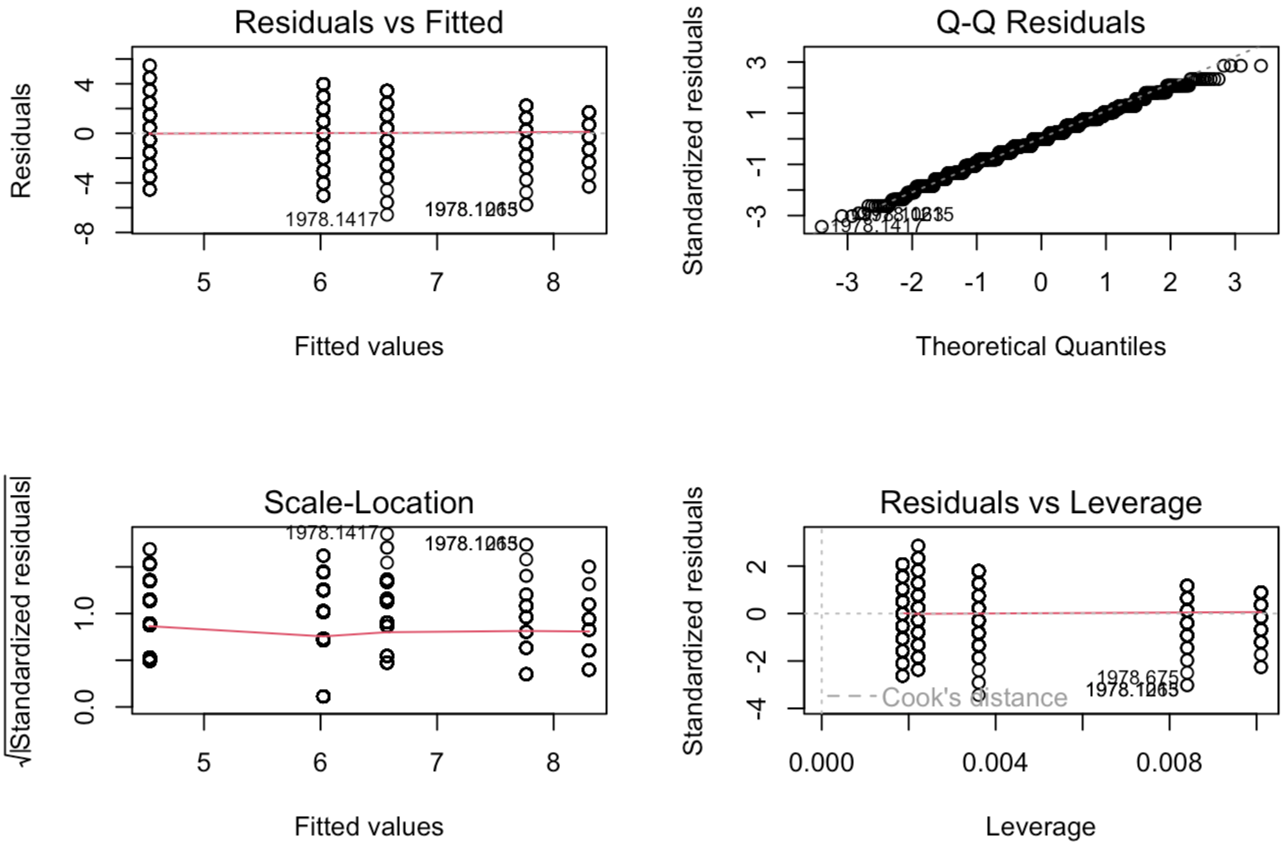
Fit model

I try an lm:

```
fit_vocab <- lm(
  formula = vocab ~ educGroup,
  data = vocab
)
```

Loading [MathJax]/jax/output/HTML-CSS/jax.js

```
par(mfrow = c(2, 2))
plot(fit_vocab)
```



Woww we see something looks really off with independence and homoscedasticity. But normality and outliera don't look too bad. It's because our data is not continous.

Normally I would do an ordinal regression now, but to report the stats I've been asked for, I'll ignore it for now.

```
summary(fit_vocab)
```

Loading [MathJax]/jax/output/HTML-CSS/jax.js

```
## 
## Call:
## lm(formula = vocab ~ educGroup, data = vocab)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.5704 -1.3030 -0.0242  1.4296  5.4656 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         4.53437    0.09025   50.24   <2e-16 ***
## educGroup12 yrs     1.48980    0.12237   12.18   <2e-16 ***
## educGroup13-15 yrs  2.03603    0.14631   13.92   <2e-16 ***
## educGroup16 yrs     3.23034    0.19752   16.35   <2e-16 ***
## educGroup>16 yrs    3.76866    0.21272   17.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.917 on 1479 degrees of freedom
##   (48 observations deleted due to missingness)
## Multiple R-squared:  0.2644, Adjusted R-squared:  0.2624 
## F-statistic: 132.9 on 4 and 1479 DF,  p-value: < 2.2e-16
```

Education group significantly predicted vocabulary performance, *F*(4, 1479) = 132.90, *p* < .001, explaining 26.2% of the variance in scores (R^2 = .262).

Compared to individuals with less than 12 years of education, vocabulary scores were significantly higher for participants with 12 years (*b* = 1.49, *SE* = 0.12, *p* < .001) and 13–15 years (*b* = 2.04, *SE* = 0.15, *p* < .001), with scores being 1.5-2 points higher. It seems that at 15+, which would approximately be the cutoff for master’s and doctorate degrees, the estimated increment in vocabulary scores jumps a bit. The model estimated that people with 16 years of education had approximately 3.23 points higher score than people with less than 12 years (**b* = 3.23, *SE* = 0.20, *p* < .001), and more than 16 years of education (*b* = 3.77, *SE* = 0.21, *p* < .001) warranted a score approximately 3.77 times higher than baseline. Vocabulary performance increased with higher levels of educational attainment.

Including nativeness

I will do the plot again:

Loading [MathJax]/jax/output/HTML-CSS/jax.js

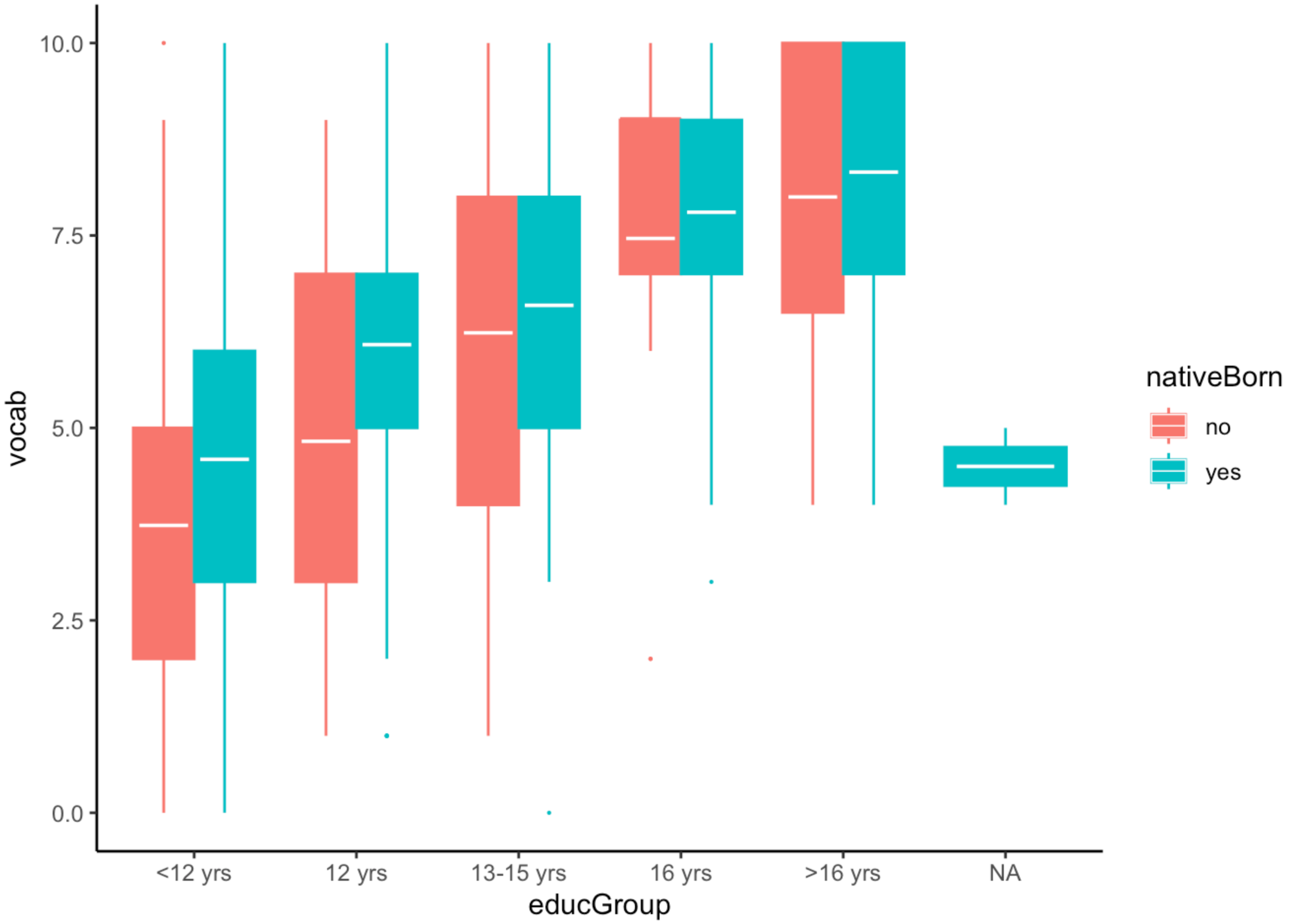
```
# for the plot I want to exclude NA
plot_data <- subset(vocab, !is.na(nativeBorn))

ggplot(
  plot_data,
  aes(
    x = educGroup,
    y = vocab,
    fill = nativeBorn,
    col  = nativeBorn
  )
) +
  geom_boxplot(
    position = position_dodge(width = 0.75),
    outlier.size = 0.1
  ) +
  stat_summary(
    fun = mean,
    geom = "crossbar",
    aes(group = nativeBorn),
    position = position_dodge(width = 0.75),
    color = "white",
    size = 0.25,
    width = 0.6
  ) +
  theme_classic()
```

```
## Warning: Removed 45 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

```
## Warning: Removed 45 rows containing non-finite outside the scale range
## (`stat_summary()`).
```

Loading [MathJax]/jax/output/HTML-CSS/jax.js



Okay, it looks like both for no and yes, the plots follow pretty similar trajectories. Here, it looks like nativeness could make a difference at the shorter durations of education, but less and less as education groups progresses. We will explore this furhther in our final model, but first we will just look at what including nativeness as a predictor alongside education group does:

Now we fit our model.

```
fit_vocab2 <- lm(
  formula = vocab ~ educGroup + nativeBorn,
  data = vocab
)

summary(fit_vocab2)
```

Loading [MathJax]/jax/output/HTML-CSS/jax.js

```
##
## Call:
## lm(formula = vocab ~ educGroup + nativeBorn, data = vocab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6170 -1.1952 -0.0604  1.3830  6.1740
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.8260     0.2150  17.793  < 2e-16 ***
## educGroup12 yrs      1.4756     0.1220  12.094  < 2e-16 ***
## educGroup13-15 yrs   2.0321     0.1457  13.946  < 2e-16 ***
## educGroup16 yrs      3.2628     0.1969  16.570  < 2e-16 ***
## educGroup>16 yrs     3.7642     0.2118  17.768  < 2e-16 ***
## nativeBornyes        0.7588     0.2093   3.626 0.000297 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.909 on 1477 degrees of freedom
##   (49 observations deleted due to missingness)
## Multiple R-squared:  0.2711, Adjusted R-squared:  0.2686
## F-statistic: 109.9 on 5 and 1477 DF,  p-value: < 2.2e-16
```

A multiple linear regression model was fitted to examine vocabulary scores as a function of educational attainment and nativity status. The overall model was statistically significant, *F*(5, 1477) = 109.90, *p* < .001, explaining 27% of the variance in vocabulary performance (R^2 = .27).

We saw again that educational groups predicted vocabulary, as in the previous model.

In addition, native-born participants demonstrated significantly higher vocabulary scores than non–native-born participants (*b* = 0.76, *SE* = 0.21, *t* = 3.63, *p* < .001), the model predicted around 0.76 better scores on average after adjusting for educational attainment.

So the model here explained slightly more variance than the one that only indcluded educational groups, making it a better model on paper - however, for the added number of predictors (1), the increase in explained variance (1%), does not seem that high.

Finally, we will fit an interaction model:

```
fit_vocab3<- lm(
  formula = vocab ~ educGroup * nativeBorn,
  data = vocab
)

summary(fit_vocab3)
```

Loading [MathJax]/jax/output/HTML-CSS/jax.js

```
##
## Call:
## lm(formula = vocab ~ educGroup * nativeBorn, data = vocab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5923 -1.1585 -0.0817  1.4077  6.2667
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.7333     0.3486  10.710  < 2e-16 ***
## educGroup12 yrs                   1.0928     0.5292   2.065   0.0391 *
## educGroup13-15 yrs                2.5020     0.5796   4.317 1.69e-05 ***
## educGroup16 yrs                   3.7282     0.6340   5.881 5.05e-09 ***
## educGroup>16 yrs                  4.2667     0.8539   4.997 6.52e-07 ***
## nativeBornyes                     0.8581     0.3608   2.378   0.0175 *
## educGroup12 yrs:nativeBornyes     0.3975     0.5438   0.731   0.4649
## educGroup13-15 yrs:nativeBornyes -0.5011     0.5989  -0.837   0.4029
## educGroup16 yrs:nativeBornyes    -0.5178     0.6671  -0.776   0.4378
## educGroup>16 yrs:nativeBornyes   -0.5355     0.8814  -0.608   0.5436
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.909 on 1473 degrees of freedom
##   (49 observations deleted due to missingness)
## Multiple R-squared:  0.2726, Adjusted R-squared:  0.2682
## F-statistic: 61.34 on 9 and 1473 DF,  p-value: < 2.2e-16
```

Here, we see that the overall model was statistically significant, *F*(9, 1477) = 61.34, $p < .001$, explaining 26.8% of the variance in vocabulary performance ($R^2 = .268$). This means this model is slightly better than the education group-only model, but worse than the one that includes both education group and nativeness as two separate predictor.

The interaction terms demonstrated no significant effect ($p > .05$), meaning that education group and nativeness did not have a strong impact on vocabulary score. When we interpret the beta estimates, we see that people with educations of 12 years or less, the beta estimates are positive ($b > 0$), meaning nativeborn speakers perform higher than non-native speakers. However, as the education durations grow, beta retains negative values ($b < 0$): This means that in high-educated groups, being a native speaker was less predictive of vocabulary scores than in lower-educated groups. Once again, this effect remains insignificant.

Altogether, while both nativeness and education group were strong predictors for vocabulary, we found no interaction between the predictors. At a quick glance, it seems the model that included both predictors separately performed best at explaining the variance. However, it was only slightly better than the model that only included education group. Altogether, since it is not desirable to include predictors that add little information to the model, we deem that the best model was:

$$VocabularyScore = a + b * EducationGroup$$

Loading [MathJax]/jax/output/HTML-CSS/jax.js