divorce_margarine - Villiam Molte Jensen

2024-10-30

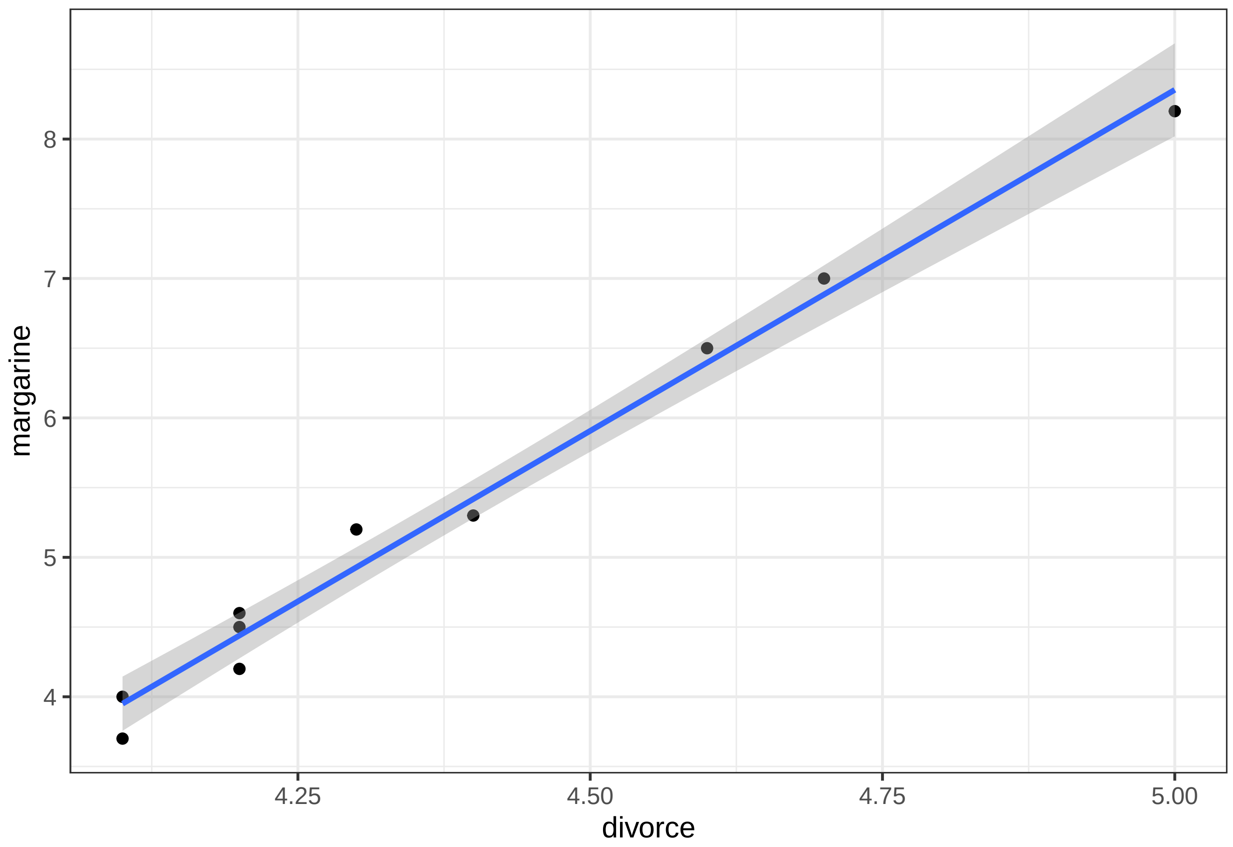
Divorce Rates and Margarine

```r
#loading in data
data("divorce_margarine")

#renaming variables
div_df <- divorce_margarine %>%
  rename(divorce = divorce_rate_maine,
         margarine = margarine_consumption_per_capita)
```

```r
#plotting correlation
div_df %>%
  ggplot(aes(x = divorce, y = margarine)) +
  geom_point() +
  geom_smooth(method = 'lm')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
#running model
model1 <- lm(margarine ~ divorce, data = div_df)

summary(model1)
```

```
##
## Call:
## lm(formula = margarine ~ divorce, data = div_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25025 -0.14422  0.05515  0.11187  0.27136
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16.1068     0.9313  -17.30 1.27e-07 ***
## divorce       4.8920     0.2122   23.05 1.33e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1893 on 8 degrees of freedom
## Multiple R-squared:  0.9852, Adjusted R-squared:  0.9833
## F-statistic: 531.5 on 1 and 8 DF,  p-value: 1.33e-08
```

There does appear to be a correlation with divorce rates and margarine consumption. However, most likely, this correlation is not causal, i.e. margarine consumption does not directly lead to higher probability of divorce. More likely, areas in which margarine consumption is higher, divorce rates also happen to be higher. One could hypothesize that social factors which lead higher margarine consumption also might lead to higher divorce rates.

vocabulary section

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

```
data("GSSvocab")
```

```
#filtering for year and excluding NAs
gss_1978 <- GSSvocab %>%
  filter(year == 1978) %>%
  na.exclude()
```

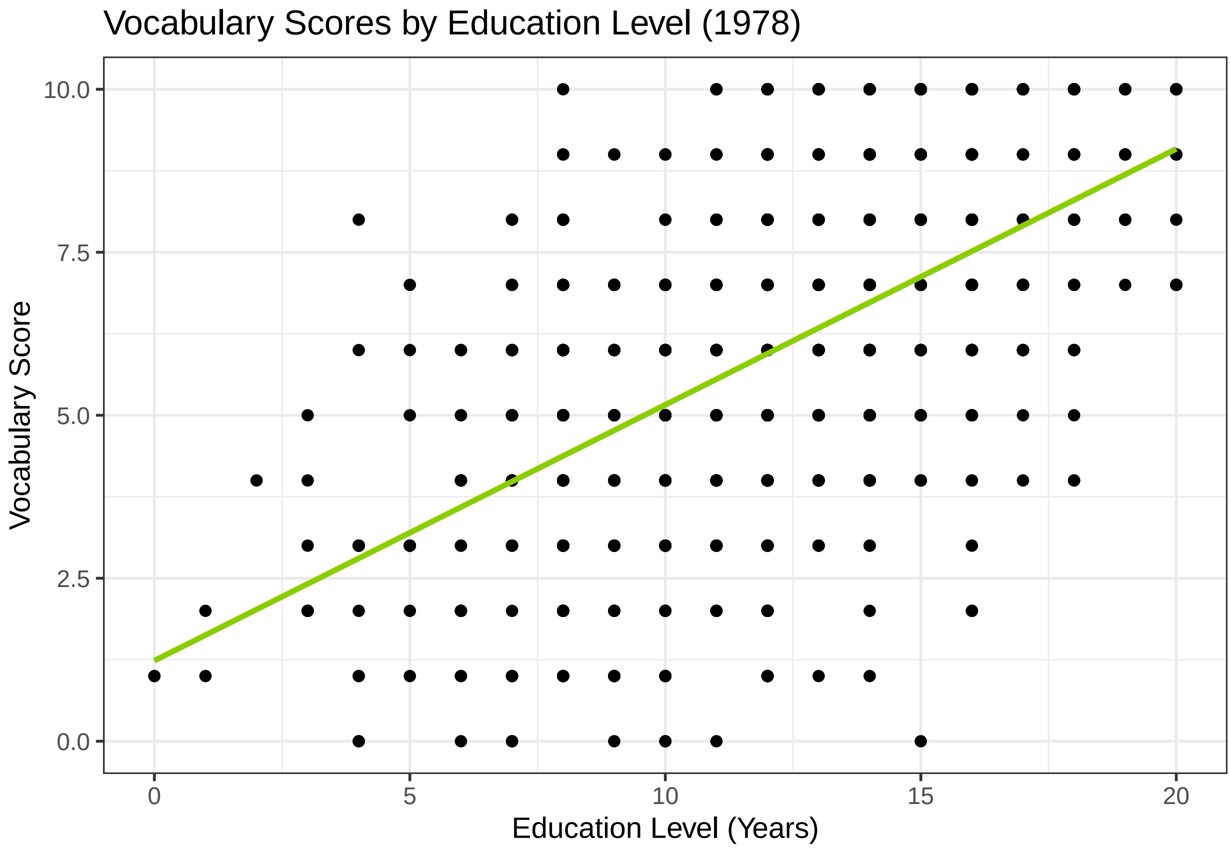
```
#creating model
model_educ <- lm(vocab ~ educ, data = gss_1978)

#summary
summary(model_educ)
```

```
##
## Call:
## lm(formula = vocab ~ educ, data = gss_1978)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1233 -1.1608  0.0542  1.0917  5.6243
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.23567    0.19957   6.192  7.7e-10 ***
## educ         0.39251    0.01606  24.443  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.885 on 1475 degrees of freedom
## Multiple R-squared:  0.2883, Adjusted R-squared:  0.2878
## F-statistic: 597.5 on 1 and 1475 DF,  p-value: < 2.2e-16
```

```r
#plotting relationship
ggplot(gss_1978, aes(x = educ, y = vocab)) +
  geom_point() +
  geom_smooth(method = "lm", color = "#8ACE00", se = FALSE) +
  labs(title = "Vocabulary Scores by Education Level (1978)",
       x = "Education Level (Years)",
       y = "Vocabulary Score")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Interpreting the model and the figure above, it seems that there is, on average, an increase in vocabulary score for each extra year in the education level.

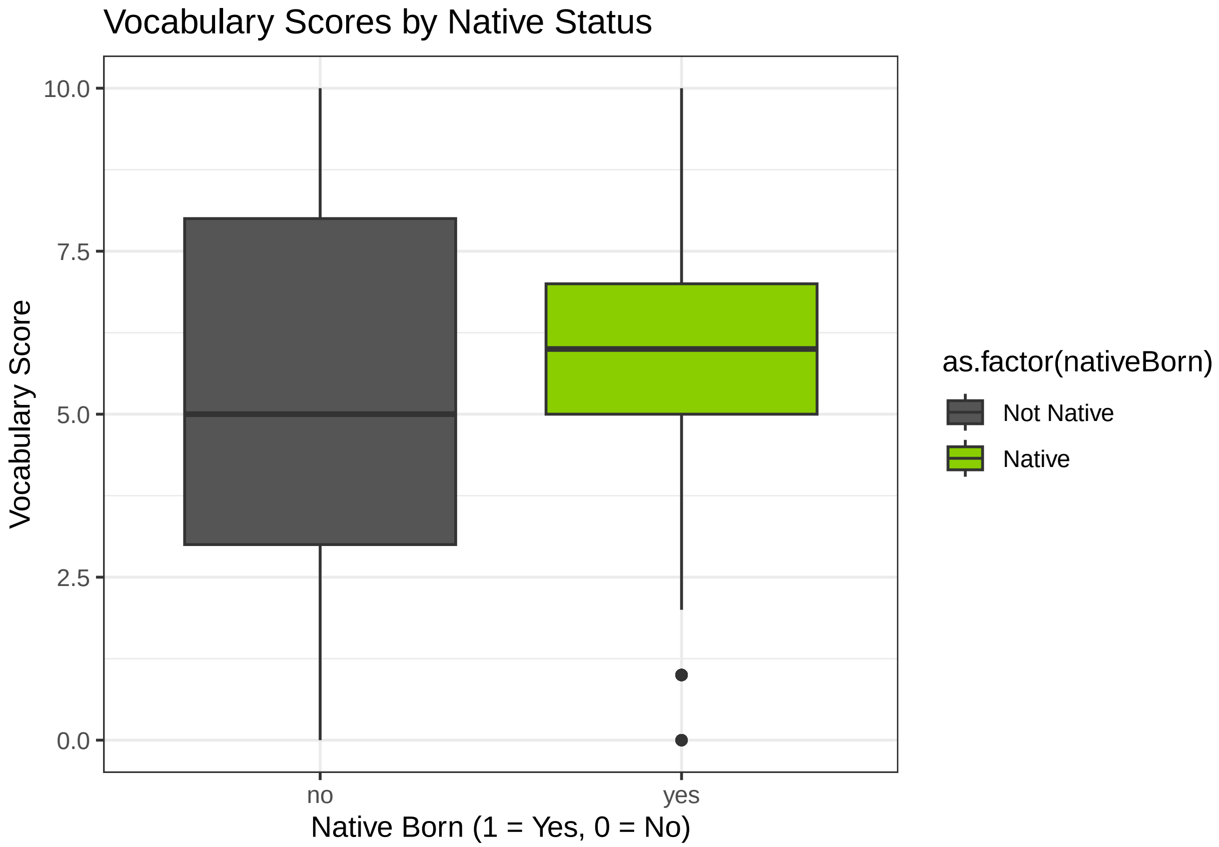
```r
#new model including nativeBorn variable
model_native <- lm(vocab ~ educ + nativeBorn, data = gss_1978)

#summary
summary(model_native)
```

```
##
## Call:
```

```
## lm(formula = vocab ~ educ + nativeBorn, data = gss_1978)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.162 -1.200  0.015  1.231  5.803
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.62803    0.27651   2.271  0.02327 *
## educ           0.39222    0.01601  24.499  < 2e-16 ***
## nativeBornyes  0.65032    0.20551   3.164  0.00159 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.879 on 1474 degrees of freedom
## Multiple R-squared:  0.2931, Adjusted R-squared:  0.2921
## F-statistic: 305.6 on 2 and 1474 DF,  p-value: < 2.2e-16
```

```r
#plotting relationship
ggplot(gss_1978, aes(x = as.factor(nativeBorn), y = vocab, fill = as.factor(nativeBorn))) +
  geom_boxplot() +
  labs(title = "Vocabulary Scores by Native Status",
       x = "Native Born (1 = Yes, 0 = No)",
       y = "Vocabulary Score") +
  scale_fill_manual(values = c("#555555", "#8ACE00"), labels = c("Not Native", "Native"))
```



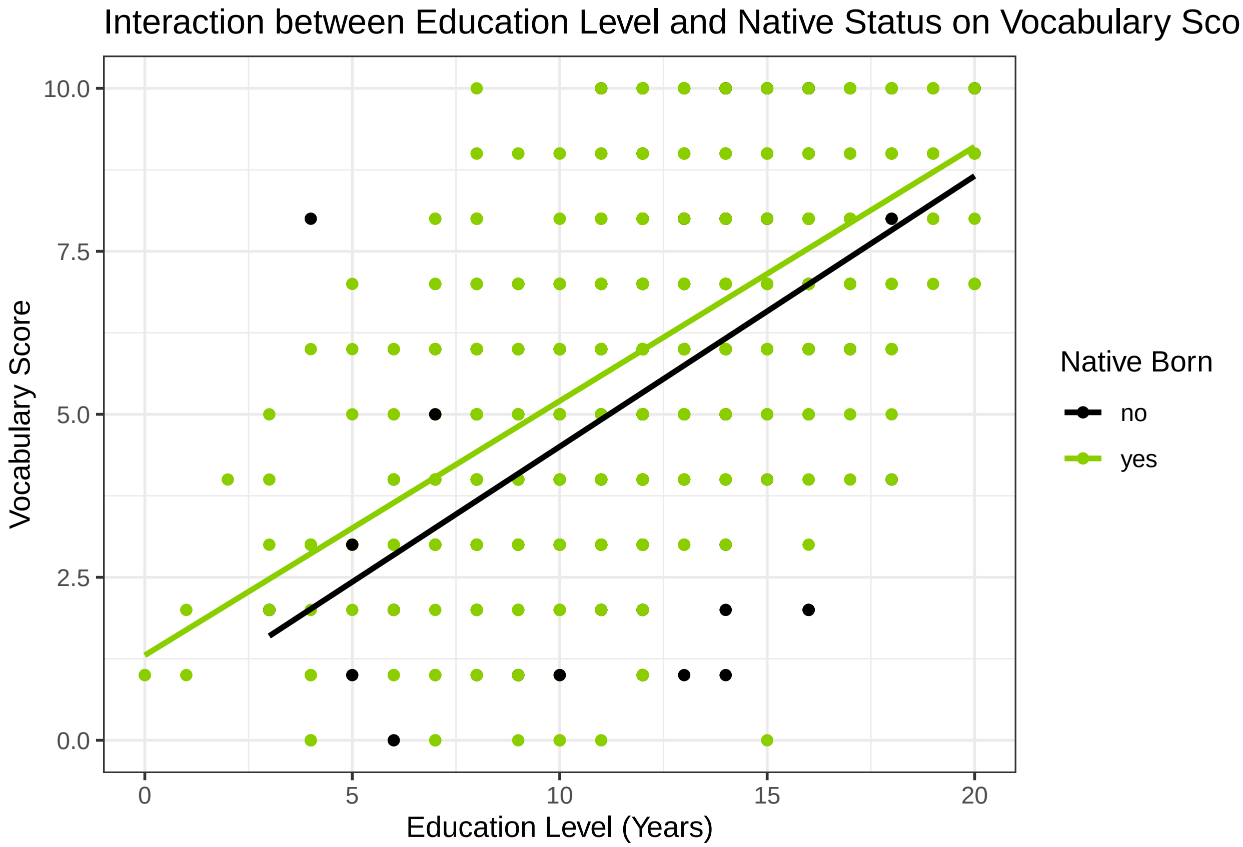
```r
# model with an interaction term
model_interaction <- lm(vocab ~ educ * nativeBorn, data = gss_1978)

#summary
summary(model_interaction)
```

```
##
## Call:
## lm(formula = vocab ~ educ * nativeBorn, data = gss_1978)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1554 -1.2049  0.0149  1.2347  5.9857
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.35394    0.68780   0.515    0.607
## educ                 0.41510    0.05496   7.553 7.45e-14 ***
## nativeBornyes        0.95000    0.71855   1.322    0.186
## educ:nativeBornyes -0.02501    0.05745  -0.435    0.663
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.88 on 1473 degrees of freedom
## Multiple R-squared:  0.2932, Adjusted R-squared:  0.2917
## F-statistic: 203.7 on 3 and 1473 DF,  p-value: < 2.2e-16
```

```r
#plotting interaction
ggplot(gss_1978, aes(x = educ, y = vocab, color = as.factor(nativeBorn))) +
  geom_point() +
  geom_smooth(method = "lm", aes(group = nativeBorn), se = FALSE) +
  labs(title = "Interaction between Education Level and Native Status on Vocabulary Scores",
       x = "Education Level (Years)",
       y = "Vocabulary Score",
       color = "Native Born") +
  scale_color_manual(values = c('black', '#8ACE00'))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The results of this model and the figure above suggest that whether being native born or not does not impact how much education level affects your vocabulary score, i.e. the positive linear relationship between these two aforementioned factors are not significantly different between native born and non-native born speakers.

```
#comparing models
summary(model_educ)$adj.r.squared
```

```
## [1] 0.2878055
```

```
summary(model_native)$adj.r.squared
```

```
## [1] 0.2921314
```

```
summary(model_interaction)$adj.r.squared
```

```
## [1] 0.2917419
```

```
#AIC values
AIC(model_educ, model_native, model_interaction)
```

```
##                   df      AIC
## model_educ         3 6068.397
## model_native       4 6060.397
## model_interaction  5 6062.207
```

Model Comparison Looking at R2 values and AIC, they all seem to have similar values. Therefore, it is hard to define one model as being vastly better than the rest. The ‘native’ model does have the highest R2 and the lowest AIC, so it seems to perform ever-so-slighty better compared to the other two models.