Assignment 2

Cultural Data Science

Glorija Stvol

November 6, 2024

Preparing the data

Load the 'divorce_margarine' dataset from the 'dslabs'package. Load the 'GSSvocab'dataset from the 'car'package.

```
#load the divorce_margarine dataset
data("divorce_margarine")
df <- divorce_margarine

# Load the GSSvocab dataset
data("GSSvocab")
df2 <- GSSvocab
```

```
#take a look at the data and summary
head(df)
```

```
##   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
```

```
summary(df)
```

```
##  divorce_rate_maine margarine_consumption_per_capita      year
##  Min.   :4.10       Min.   :3.700                    Min.   :2000
##  1st Qu.:4.20       1st Qu.:4.275                    1st Qu.:2002
##  Median :4.25       Median :4.900                    Median :2004
##  Mean   :4.38       Mean   :5.320                    Mean   :2004
##  3rd Qu.:4.55       3rd Qu.:6.200                    3rd Qu.:2007
##  Max.   :5.00       Max.   :8.200                    Max.   :2009
```

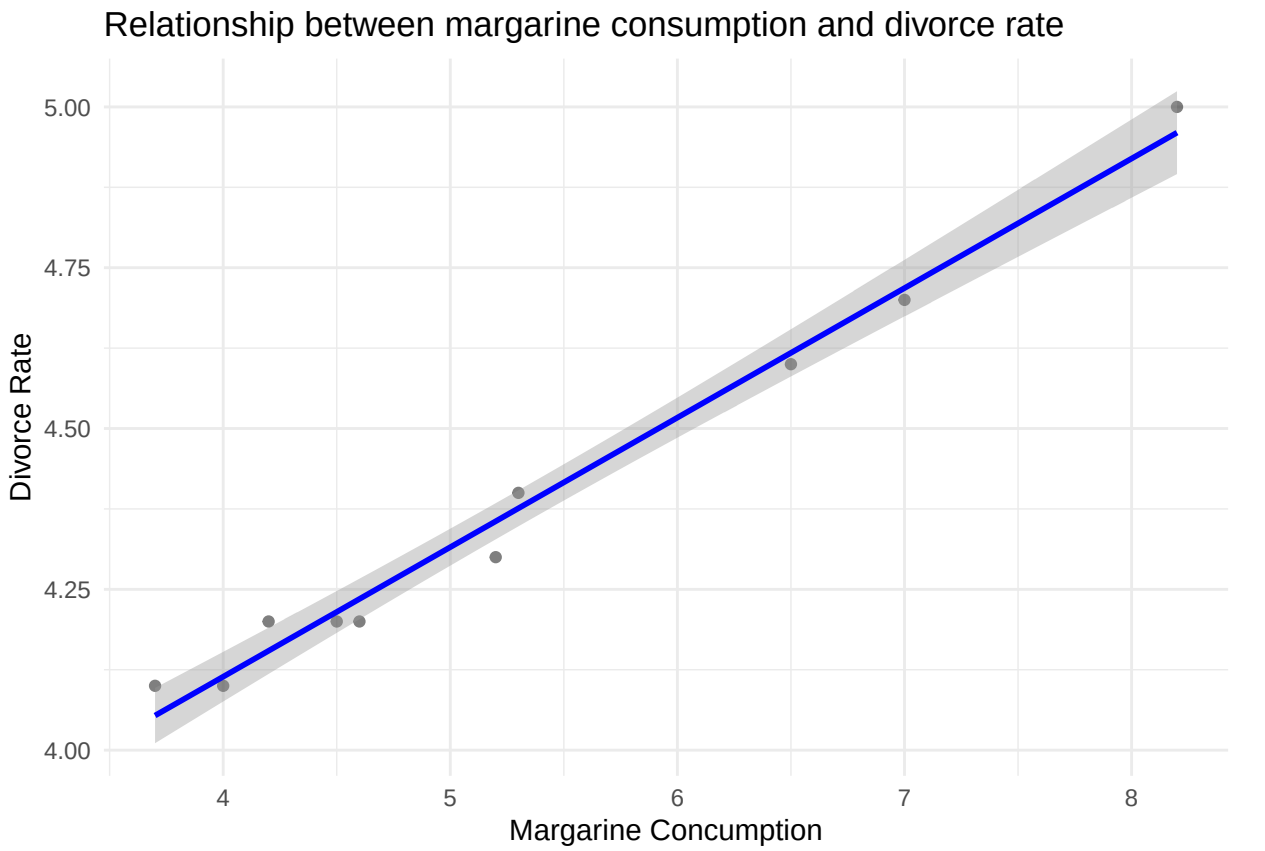
Part 1

Investigate the correlation between margarine consumption and divorce rates in Maine. Would an increase in the preference for margarine lead to skyrocketing divorce rates?

```
# Visualize relationship
ggplot(df, aes(x = margarine_consumption_per_capita, y = divorce_rate_maine)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", color = "blue") +
```

```
  labs(title = "Relationship between margarine consumption and divorce rate",
       x = "Margarine Concumption",
       y = "Divorce Rate") +
  theme_minimal()
```

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




```
#correlation test
cor.test(df$margarine_consumption_per_capita, df$divorce_rate_maine , method = c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$margarine_consumption_per_capita and df$divorce_rate_maine
## t = 23.055, df = 8, p-value = 1.33e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9676666 0.9983038
## sample estimates:
##       cor
## 0.9925585
```

```
#making a linear model
lm(df$divorce_rate_maine ~ df$margarine_consumption_per_capita)
```

```
##
## Call:
## lm(formula = df$divorce_rate_maine ~ df$margarine_consumption_per_capita)
```

```
##
## Coefficients:
##                          (Intercept)  df$margarine_consumption_per_capita
##                               3.3086                               0.2014
```

Investigate the correlation between margarine consumption and divorce rates in Maine. Would an increase in the preference for margarine lead to skyrocketing divorce rates?

Answer: Based on visual inspection of the data as well as a correlation test (cor.test), the results indicate that there is a very strong positive correlation between margarine consumption per capita and divorce rates in Maine with a correlation coefficient equal to 0.9925585. This suggests that as margarine consumption increases, the divorce rate tends to increase as well. Even though, correlation is statistically significant (p-value = 1.33e-08), without the knowledge of the theoretical background of the data set we cannot assume any causality between the variables. The observed correlation could be influenced by other factors or variables that are not accounted for in this analysis. To make any causality conclusions it is essential to explore the context with further investigation to understand the present relationships better.

Part 2

This dataset contains people's scores on an English vocabulary test and includes demographic information. Filter for the year 1978 and remove rows with missing values (the function na.exclude()is one way to do this–check out the documentation!).

```
head(df2)
```

```
##        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
```

```
summary(df2)
```

```
##       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
```

```
# Filter for rows where year is 1978
df2_filtered <- df2 %>% filter(year == 1978)

head(df2_filtered)
```

```
##        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
```

```
summary(df2_filtered) #also checking for NAs
```

```
##       year         gender    nativeBorn   ageGroup       educGroup
##  1978   :1532   female:889   no  :  92   18-29:407   <12 yrs  :485
##  1982   :   0   male  :643   yes :1438   30-39:334   12 yrs   :541
##  1984   :   0                NA's:   2   40-49:217   13-15 yrs:280
##  1987   :   0                            50-59:228   16 yrs   :119
##  1988   :   0                            60+  :339   >16 yrs  :101
##  1989   :   0                            NA's :  7   NA's     :  6
##  (Other):   0
##      vocab             age             educ
##  Min.   : 0.000   Min.   :18.00   Min.   : 0.00
##  1st Qu.: 5.000   1st Qu.:29.00   1st Qu.:10.00
##  Median : 6.000   Median :40.00   Median :12.00
##  Mean   : 5.963   Mean   :44.01   Mean   :11.92
##  3rd Qu.: 7.000   3rd Qu.:58.00   3rd Qu.:14.00
##  Max.   :10.000   Max.   :89.00   Max.   :20.00
##  NA's   :46       NA's   :7       NA's   :6
```

```
# Remove rows with NAs using na.exclude
df2_filtered <- na.exclude(df2_filtered)

# Check for NAs with summary
summary(df2_filtered)
```

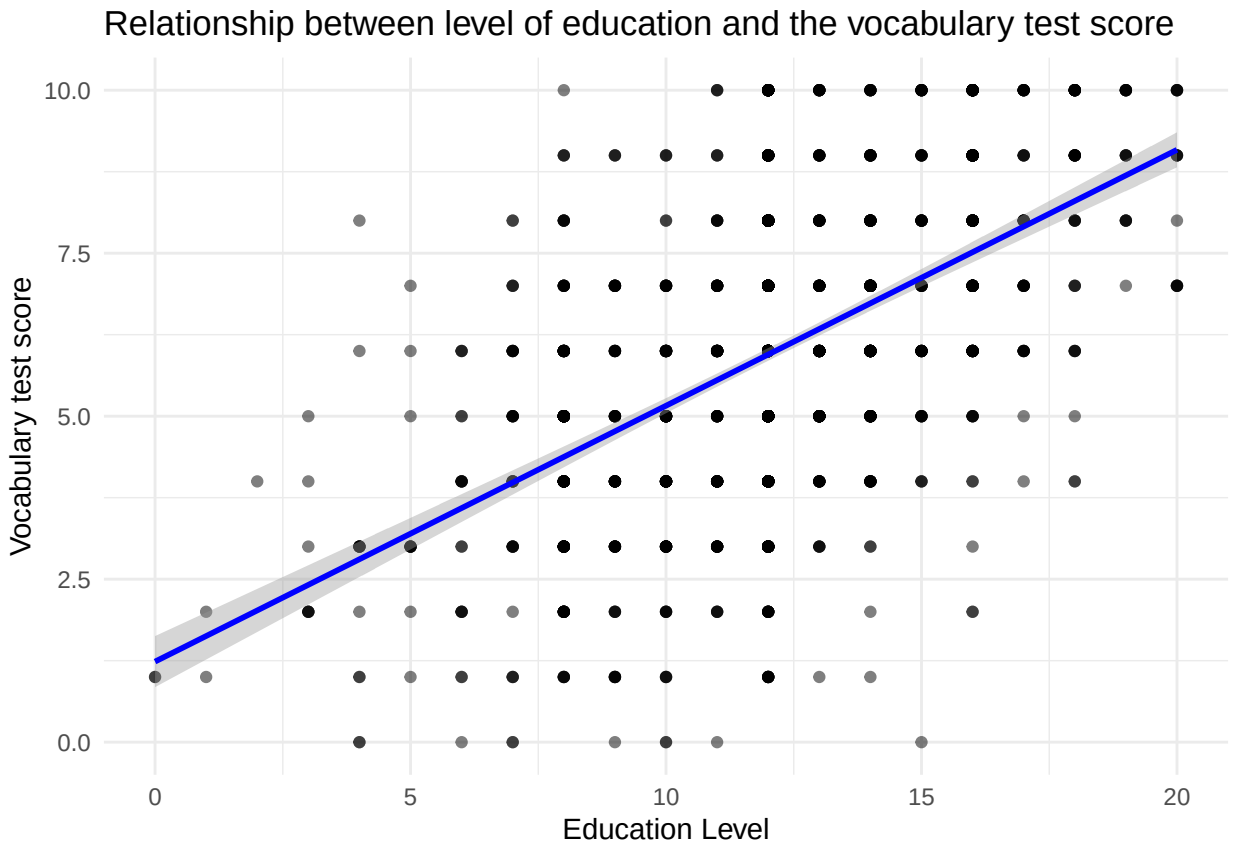
```
##       year         gender    nativeBorn  ageGroup       educGroup
##  1978   :1477   female:854   no :  89   18-29:401   <12 yrs  :449
##  1982   :   0   male  :623   yes:1388   30-39:330   12 yrs   :534
##  1984   :   0                           40-49:208   13-15 yrs:276
##  1987   :   0                           50-59:220   16 yrs   :119
##  1988   :   0                           60+  :318   >16 yrs  : 99
##  1989   :   0
##  (Other):   0
##      vocab             age             educ
##  Min.   : 0.000   Min.   :18.00   Min.   : 0.00
##  1st Qu.: 5.000   1st Qu.:29.00   1st Qu.:11.00
##  Median : 6.000   Median :40.00   Median :12.00
##  Mean   : 5.964   Mean   :43.58   Mean   :12.05
##  3rd Qu.: 7.000   3rd Qu.:57.00   3rd Qu.:14.00
##  Max.   :10.000   Max.   :89.00   Max.   :20.00
##
```

2.1

Is a person's score on the vocabulary test ('vocab') significantly impacted by their level of education ('educ')? Visualize the relationship in a plot and build a model. Briefly explain the results.

```r
# Visualize relationship
ggplot(df2_filtered, aes(x = educ, y = vocab)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "Relationship between level of education and the vocabulary test score",
       x = "Education Level",
       y = "Vocabulary test score") +
  theme_minimal()
```

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




```r
#make a model
m1 <- lm(vocab~educ, df2_filtered)
summary(m1)
```

```
##
## Call:
## lm(formula = vocab ~ educ, data = df2_filtered)
##
## 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
```

Answer:

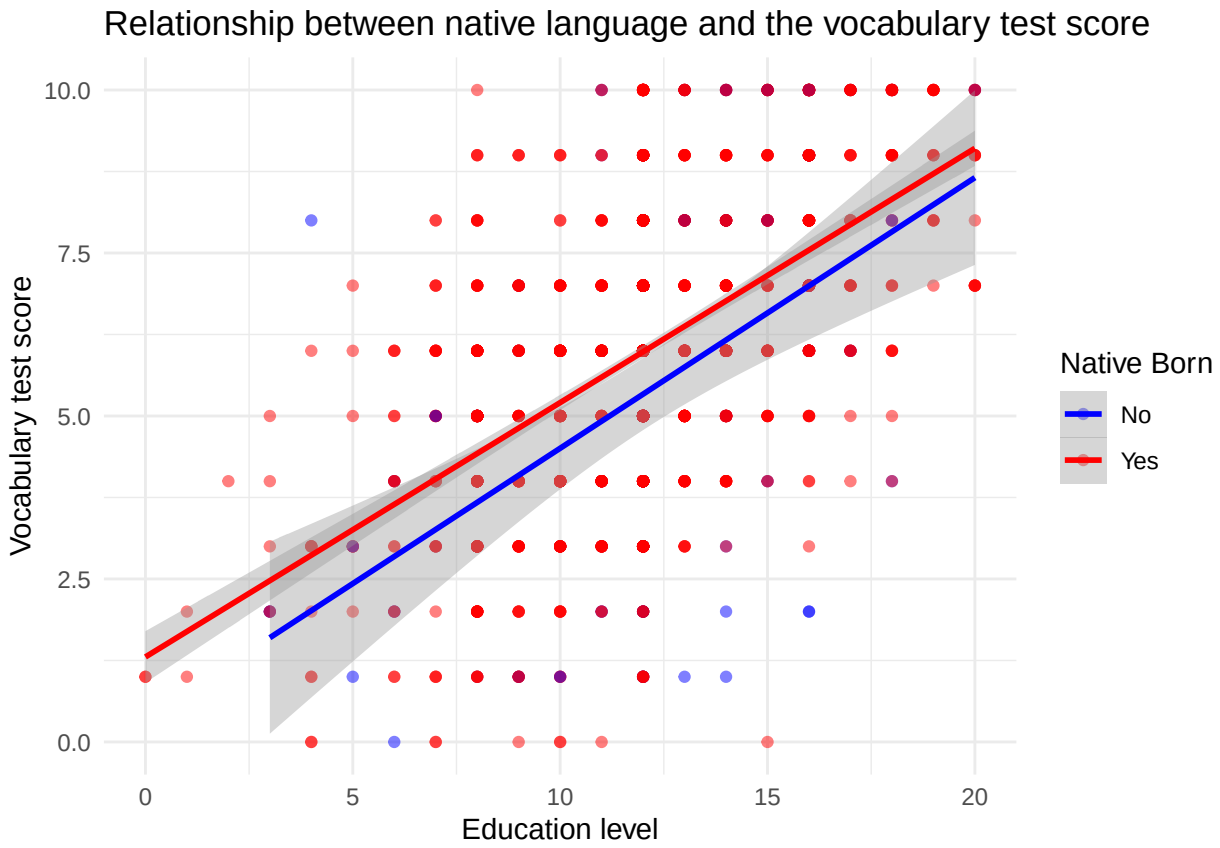
This model indicates a positive and significant association between education level and vocabulary scores. The low p-value implies that the effect of education on vocabulary score is statistically significant. Nevertheless education alone explains only partly the variation in vocabulary, as the R-squared value of 0.288 suggests that about 28.8% of the variability in vocabulary scores can be explained by education level.

2.2

Whether a person is the native of an English-speaking country ('nativeBorn') could potentially have an impact on the size of their vocabulary. Visualize the relationship and add the predictor to the model. Briefly explain the results.

```
#Visualize the relationship
ggplot(df2_filtered, aes(x = educ, y = vocab, color = nativeBorn)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm") +
  labs(
    title = "Relationship between native language and the vocabulary test score",
    x = "Education level",
    y = "Vocabulary test score",
    color = "Native Born",
    linetype = "Native Born"
  ) +
  scale_color_manual(values = c("blue", "red"), labels = c("No", "Yes")) +
  theme_minimal()
```

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




```
#make a model and add a predictor
m2 <- lm(vocab ~ educ + nativeBorn, df2_filtered)
summary(m2)
```

```
##
## Call:
## lm(formula = vocab ~ educ + nativeBorn, data = df2_filtered)
##
## 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
```

Answer:

The estimate for nativeBornyes (0.65) suggests that being a native English speaker is associated with an additional 0.65 points in vocabulary score, holding education level consistent with the previous model. This

effect is statistically significant (p = 0.00159). The R-squared value slightly increased to 0.293, indicating that adding nativeBorn improved the model's explanatory power, though the increase is small. The results suggest that higher education levels predict higher vocabulary scores and being a native of an English-speaking country adds a modest but significant increase in vocabulary scores.

2.3

Does a person's level of education depend on whether they are a native of the country? Do you think it makes sense to add the relationship as an interaction term? Try creating the model and briefly explain the results.

```
#make a model with interaction
m3 <- lm(vocab ~ educ * nativeBorn, df2_filtered)
summary(m3)
```

```
##
## Call:
## lm(formula = vocab ~ educ * nativeBorn, data = df2_filtered)
##
## 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
```

Answer:

This model indicates that native-born status does not significantly impact baseline vocabulary scores when education is zero, and the interaction term between education and native-born status is also not significant (p = 0.663). This indicates that the relationship between education and vocabulary score is similar for both native and non-native English speakers. Adding the interaction term does not improve the model's explanatory power, as reflected in the very small change in R-squared (remaining around 0.293). Thus, the simpler model without the interaction term is both sufficient and more interpretable, with education having a consistent positive effect on vocabulary regardless of native-born status.

2.4

Which model performs best?

```
#compare models
anova(m1, m2)
```

```
## Analysis of Variance Table
##
## Model 1: vocab ~ educ
## Model 2: vocab ~ educ + nativeBorn
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
```

```
## 1   1475 5241.8
## 2   1474 5206.5  1    35.371 10.014 0.001585 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m2, m3)
```

```
## Analysis of Variance Table
##
## Model 1: vocab ~ educ + nativeBorn
## Model 2: vocab ~ educ * nativeBorn
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1474 5206.5
## 2   1473 5205.8  1   0.66952 0.1894 0.6634
```

```
anova(m1, m3)
```

```
## Analysis of Variance Table
##
## Model 1: vocab ~ educ
## Model 2: vocab ~ educ * nativeBorn
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1   1475 5241.8
## 2   1473 5205.8  2     36.04 5.0989 0.006212 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# ANOVA to compare models
anova(m1, m2, m3)
```

```
## Analysis of Variance Table
##
## Model 1: vocab ~ educ
## Model 2: vocab ~ educ + nativeBorn
## Model 3: vocab ~ educ * nativeBorn
##   Res.Df    RSS Df Sum of Sq       F  Pr(>F)
## 1   1475 5241.8
## 2   1474 5206.5  1    35.371 10.0083 0.00159 **
## 3   1473 5205.8  1     0.670  0.1894 0.66344
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# AIC and BIC for model comparison
AIC(m1, m2, m3)
```

```
##    df      AIC
## m1  3 6068.397
## m2  4 6060.397
## m3  5 6062.207
```

```
BIC(m1, m2, m3)
```

```
##    df      BIC
## m1  3 6084.291
## m2  4 6081.588
## m3  5 6088.696
```

Answer:

The ANOVA test indicates that Model 2 (with educ + nativeBorn) significantly improves the fit over Model 1 (with only educ as a preditor), with a p-value of 0.00159. However, adding the interaction term in Model 3 does not provide a significant improvement over Model 2 (p = 0.66344). Model 2 has the lowest AIC (6060.4) and BIC (6081.6) values, suggesting it has the best balance between model fit and complexity. Model 3 has slightly higher AIC and BIC values, indicating that the added complexity of the interaction term does not improve model performance. Overall, Model 2 (vocab ~ educ + nativeBorn) performs best, as it significantly improves fit over the baseline model (m1) and has the lowest AIC and BIC scores without unnecessary complexity.