Assignment 2

P.H.W

2024-10-30

R Markdown

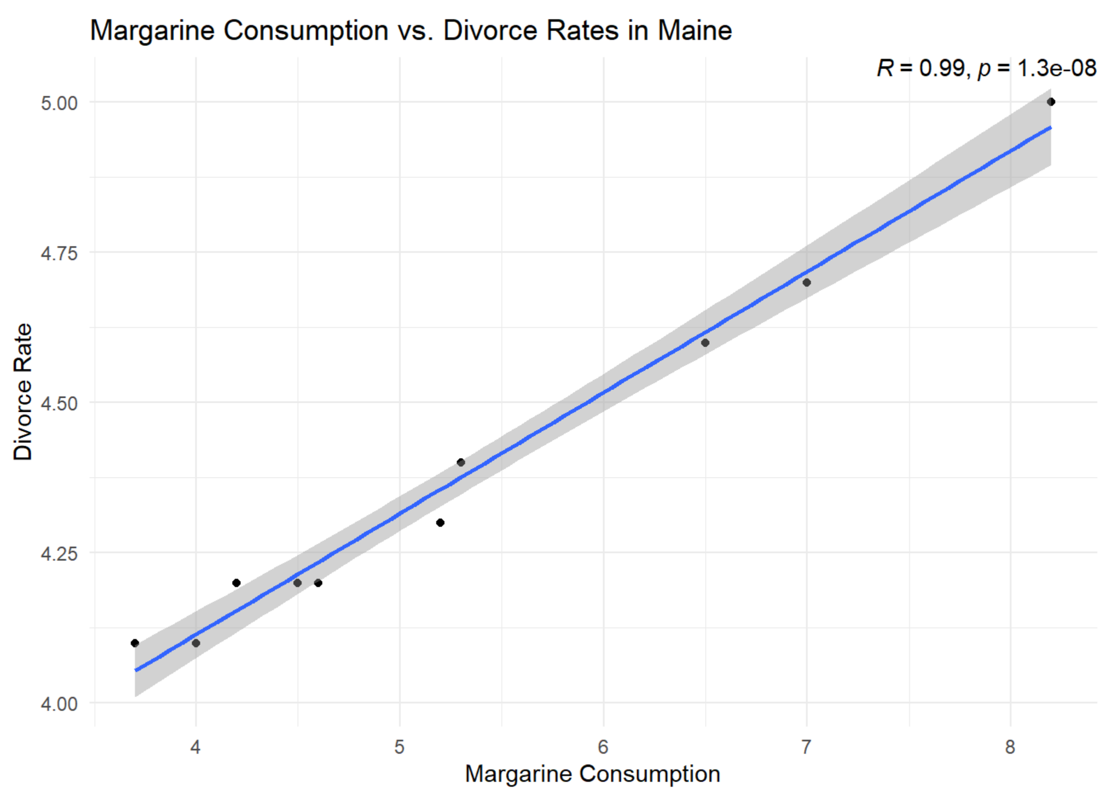
Part 1

Load the 'divorce_margarine'dataset from the 'dslabs'package. Investigate the correlation between margarine consumption and divorce rates in Maine. Would an increase in the preference for margarine lead to skyrocketing divorce rates

```
data(divorce_margarine)

divorce_margarine %>%
  ggplot(aes(x = margarine_consumption_per_capita, y = divorce_rate_maine)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(x = "Margarine Consumption", y = "Divorce Rate") +
  theme_minimal() +
  stat_cor(method = "pearson", label.x = Inf, label.y = Inf, hjust = 1, vjust = 1)+
  ggtitle("Margarine Consumption vs. Divorce Rates in Maine")
```

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




```
# We see a high positive correlation between margarine consumption and divorce rates in Maine: R = 0.99, p = 1.3e
-08
```

Part 2

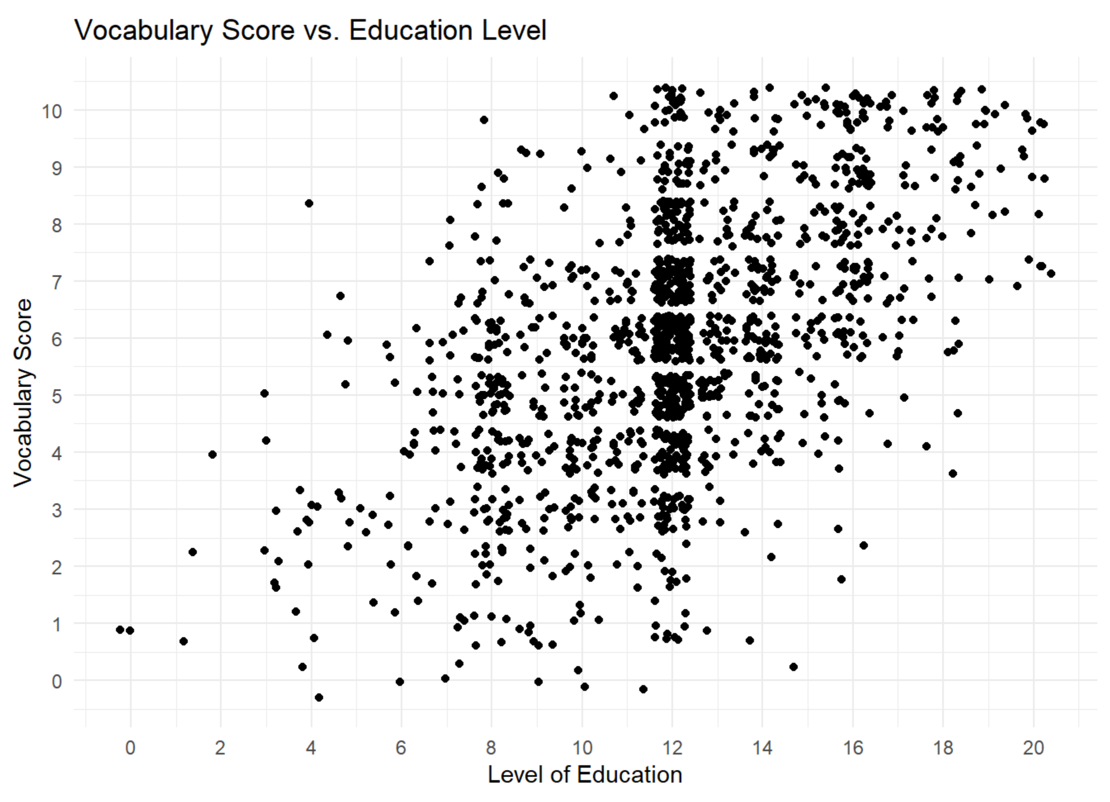
Load the 'GSSvocab'dataset from the 'car'package. This dataset contains people's scores on an Englishvocabulary test and includes demographic information. Filter for the year 1978 and remove rows with missing values (the function na.exclude()is one way to do this–check out the documentation!).

```
data(GSSvocab)


filtered <- GSSvocab %>% filter(year==1978) %>% na.exclude()
```

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 theresults.

```
filtered %>% ggplot(aes(educ,vocab)) +
  geom_point(position = 'jitter') +
  scale_y_continuous(breaks = seq(0, max(filtered$vocab, na.rm = TRUE), by = 1)) +
  scale_x_continuous(breaks = seq(0, max(filtered$educ, na.rm = TRUE), by = 2))+
  labs(x = "Level of Education", y = "Vocabulary Score") +
  theme_minimal()+
  ggtitle("Vocabulary Score vs. Education Level")
```



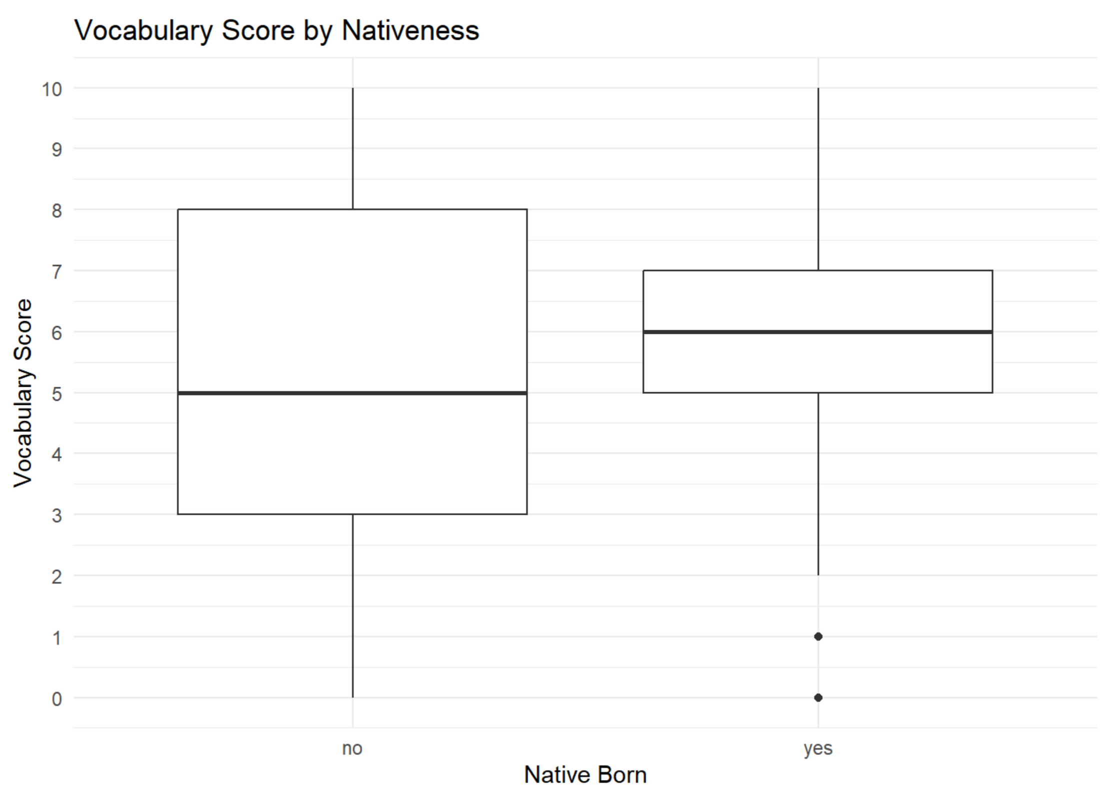

```
m1 <- lm(vocab~educ, filtered)
summary(m1)
```

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

```
# We see from the model summary that education level significantly impacts the vocabulary test score,
# and that, on average, each additional year/level of education is associated with an increase of 0.39 points in
the vocabulary score.
```

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.

```
filtered %>% ggplot(aes(x = nativeBorn, y = vocab)) +
  geom_boxplot() +
  scale_y_continuous(breaks = seq(0, max(filtered$vocab, na.rm = TRUE), by = 1))+
  labs(x = "Native Born", y = "Vocabulary Score") +
  theme_minimal()+
  ggtitle("Vocabulary Score by Nativeness")
```



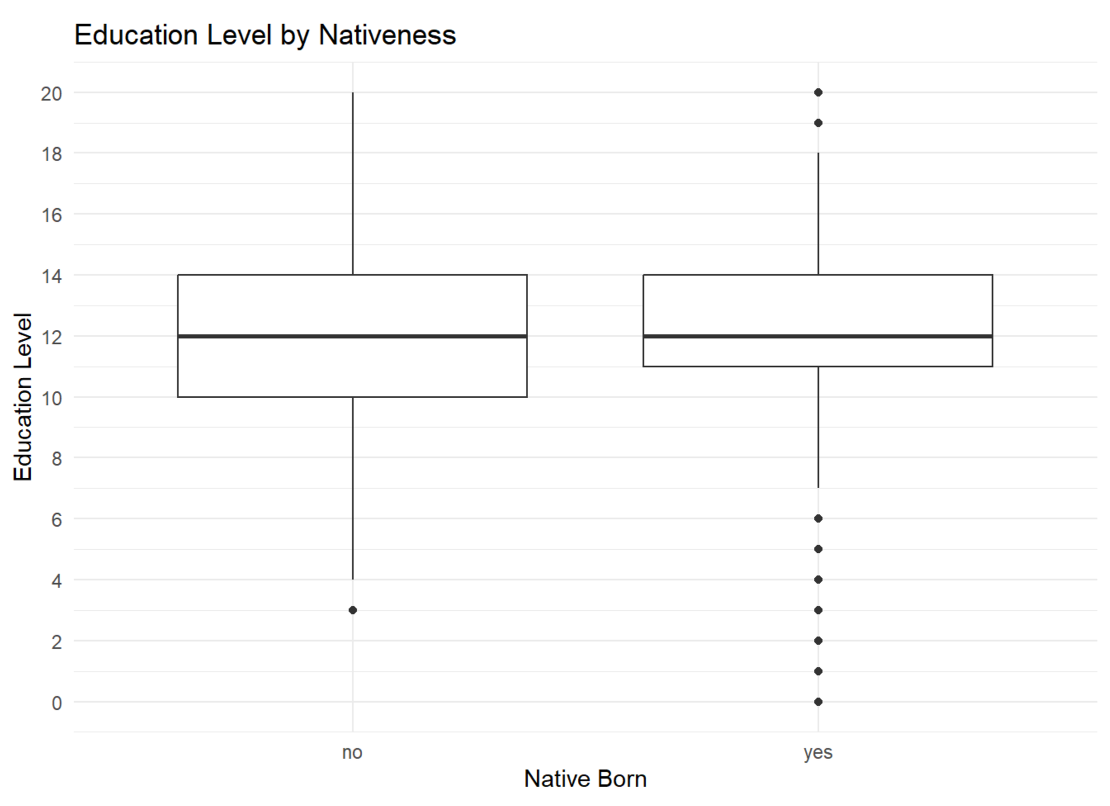

```
m2 <- lm(vocab~educ + nativeBorn, filtered)
summary(m2)
```

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

```
# From the model summary we see that the education coefficient remains consistent with the previous model,
# and that, on average, being a native-born in an English-speaking country is associated with an increase of 0.65
points in vocabulary score.
```

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

```
ggplot(data = filtered, aes(x = as.factor(nativeBorn), y = educ)) +
  geom_boxplot() +
  labs(x = "Native Born", y = "Education Level") +
  scale_y_continuous(breaks = seq(0, max(filtered$educ, na.rm = TRUE), by = 2))+
  theme_minimal() +
  ggtitle("Education Level by Nativeness")
```




```
m3 <- lm(vocab~educ*nativeBorn, filtered)
summary(m3)
```

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

```
# We see that the interaction term is not significant, which suggests that the impact of education on vocabulary
does not depend on nativeness.
# We can also see that education level is the only good predictor of vocabulary score in this model.
```

Which model performs best?

```
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(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
```

```
# In the ANOVA We see that adding nativeBorn to the model significantly improves the fit,
# and adding the interaction term (educ:nativeBorn) does not significantly improve the model fit.

# Besides significantly improving the fit, Model 2: vocab ~ educ + nativeBorn, also has the lowest AIC and BIC sc
ores.
```