Assignment 02

Réka Forgó

2025-10-31

Assignment 02

Correlation and Regression

Part 1

Use the 'divorce_margarine' dataset from the 'dslabs' package to explore the relationship between margarine consumption and divorce rates in Maine. Visualise the data, make a correlation test. Report the correlation coefficient and p-value in APA format, and briefly discuss the results in a practical way.

```
pacman::p_load(ggplot2, dplyr, tidyr, dslabs, car, pastecs, broom, performance, see)
```

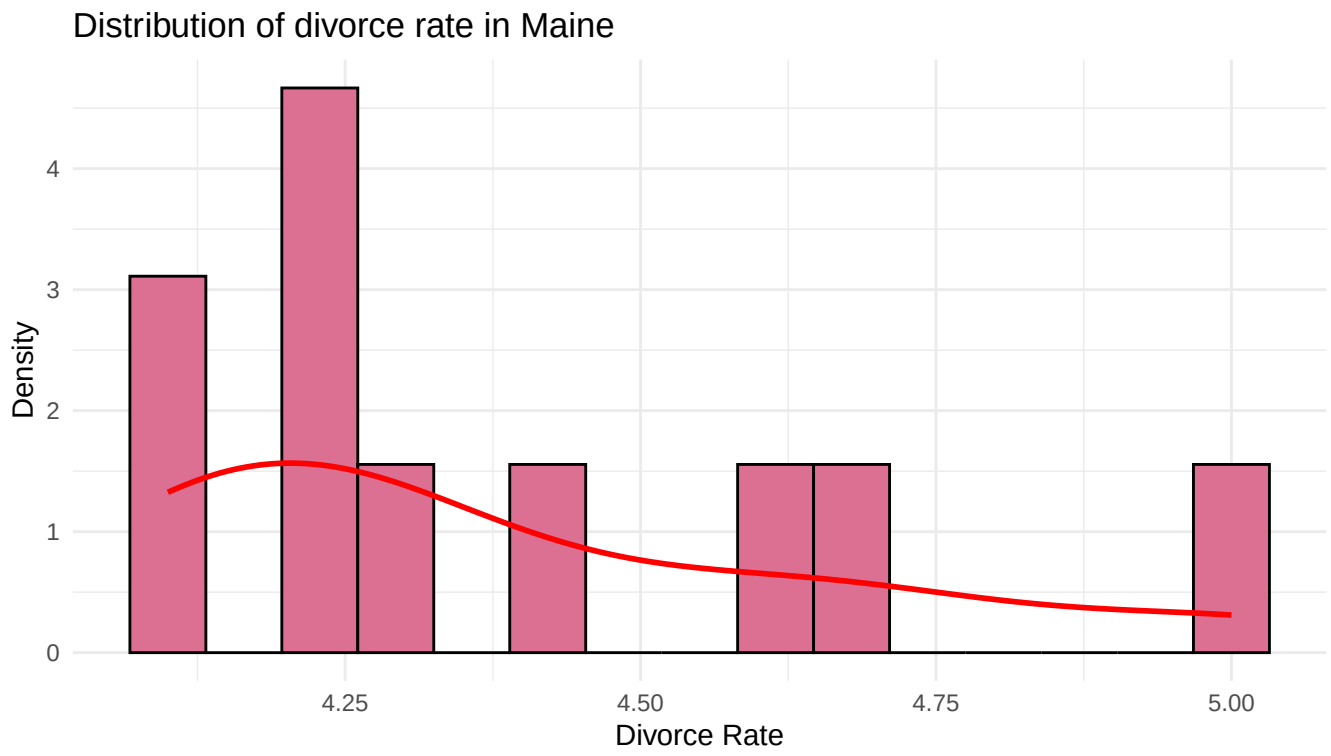
```
df <- as.data.frame(divorce_margarine)
head(df)
```

```
  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
```

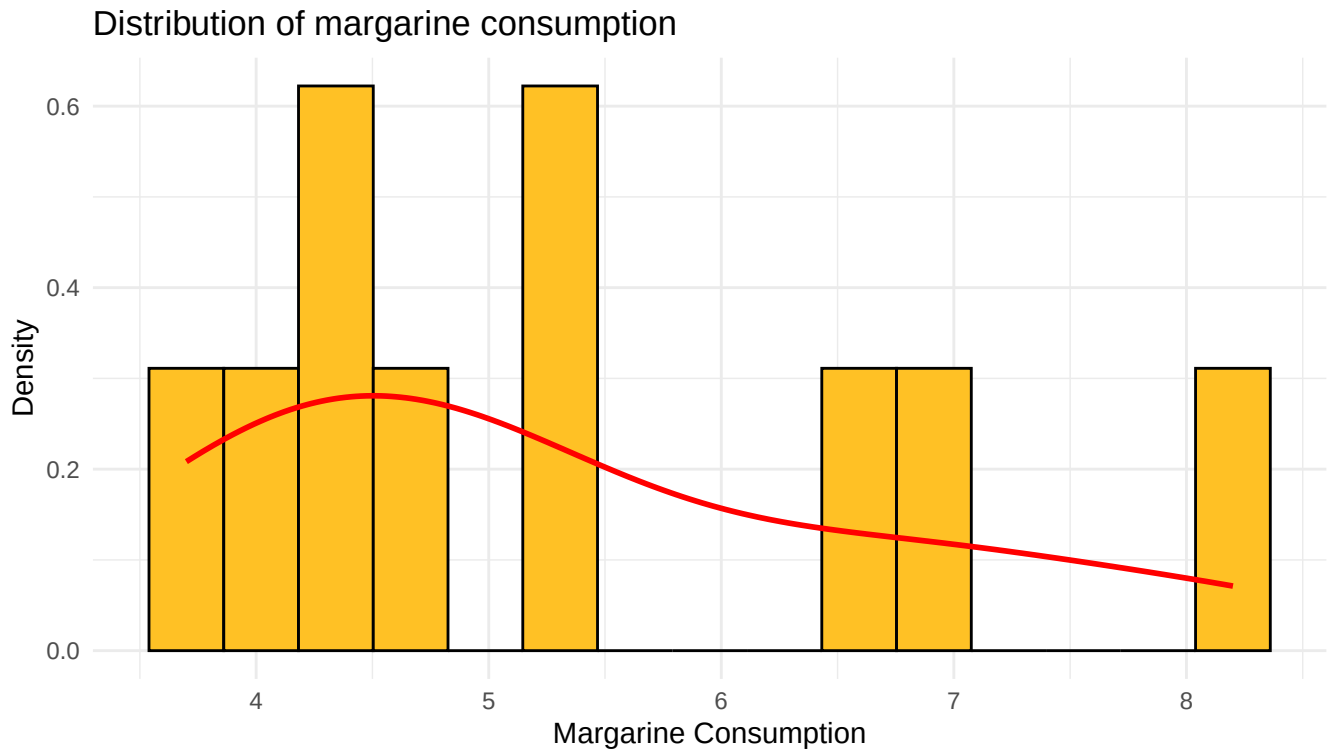
The dataset:

Columns: year: year, divorce_rate: Divorce rate in Maine(per 1000 people), margarine_consumption: Margarine consumption per capita in US(lbs). The dataset contains datapoints for 10 different years, from 2000 to 2009.

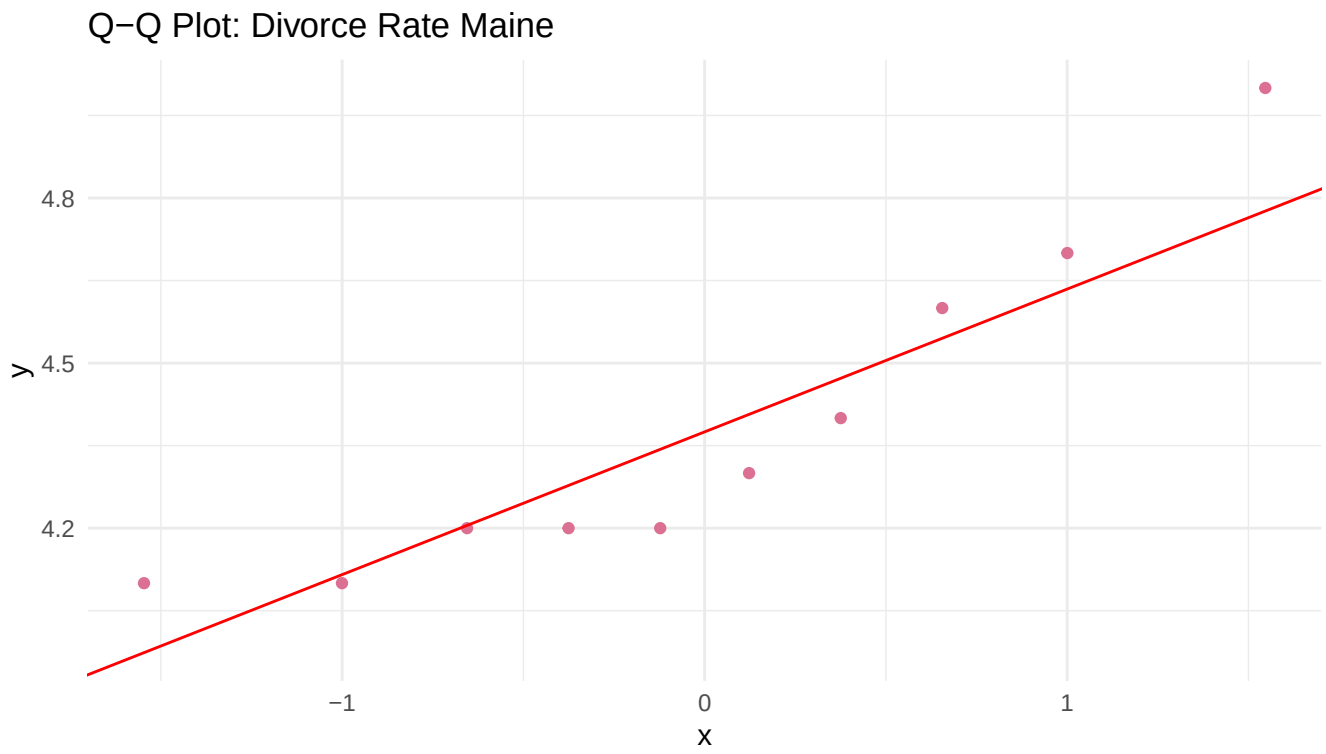
```
#normality checks

ggplot(df, aes(x = divorce_rate_maine)) +
  geom_histogram(aes(y = ..density..), bins = 15,
                 fill = "palevioletred", color = "black") +
  geom_density(color = "red", linewidth = 1) +
  labs(title = "Distribution of divorce rate in Maine",
       x = "Divorce Rate", y = "Density") +
  theme_minimal()
```



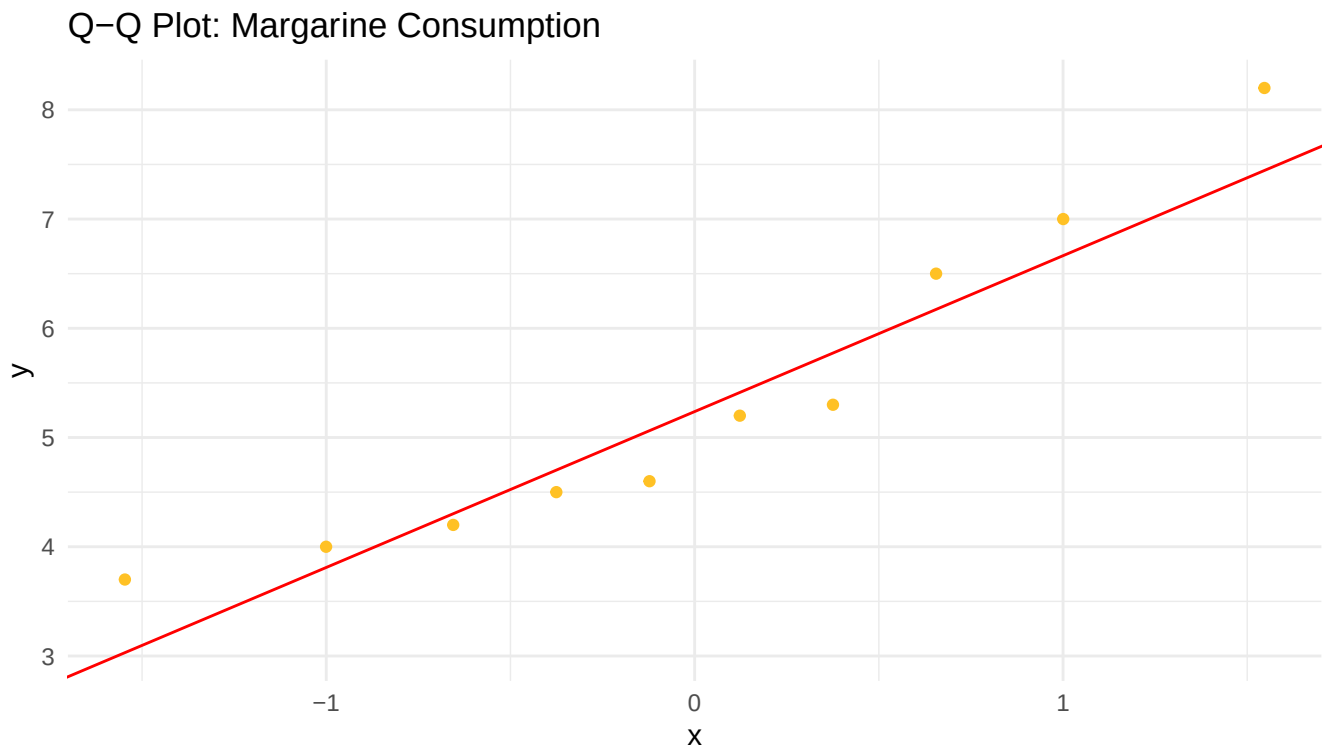

```
ggplot(df, aes(x = margarine_consumption_per_capita)) +
  geom_histogram(aes(y = ..density..), bins = 15,
                 fill = "goldenrod1", color = "black") +
  geom_density(color = "red", linewidth = 1) +
  labs(title = "Distribution of margarine consumption",
       x = "Margarine Consumption", y = "Density") +
  theme_minimal()
```




```
ggplot(df, aes(sample = divorce_rate_maine)) +
  stat_qq(color = "palevioletred") +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot: Divorce Rate Maine") +
  theme_minimal()
```




```
ggplot(df, aes(sample = margarine_consumption_per_capita)) +
  stat_qq(color = "goldenrod1") +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot: Margarine Consumption") +
  theme_minimal()
```



We can see that we have very few data points, so the histograms are not very informative. The Q-Q plots show that the data points are roughly following the diagonal line, which suggests that the data is approximately normally distributed.

```
shapiro.test(df$divorce_rate_maine)
```

```
    Shapiro-Wilk normality test

data:  df$divorce_rate_maine
W = 0.86135, p-value = 0.07916
```

```
shapiro.test(df$margarine_consumption_per_capita)
```

```
    Shapiro-Wilk normality test

data:  df$margarine_consumption_per_capita
W = 0.90531, p-value = 0.2503
```

```
round(pastecs::stat.desc(cbind(df$divorce_rate_maine,
                               df$margarine_consumption_per_capita),
                         basic = FALSE, norm = TRUE), digits = 2)
```

```
                V1    V2
median        4.25  4.90
mean          4.38  5.32
SE.mean       0.09  0.46
CI.mean.0.95  0.21  1.05
var           0.09  2.15
std.dev       0.30  1.47
```

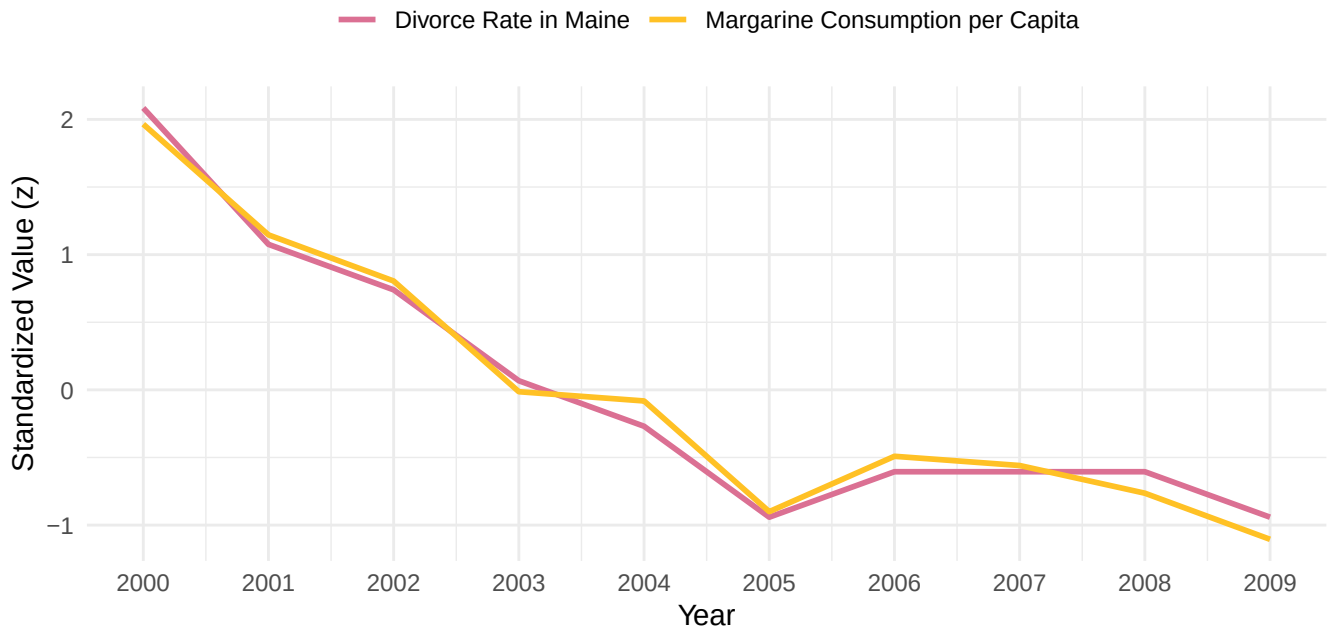
```
coef.var      0.07  0.28
skewness      0.84  0.68
skew.2SE      0.61  0.49
kurtosis     -0.75 -1.03
kurt.2SE     -0.28 -0.39
normtest.W    0.86  0.91
normtest.p    0.08  0.25
```

For both the divorce rate and margarine consumption, the Shapiro-Wilk test yields p values greater than .05, which means we can treat the data as approximately normally distributed. Both W statistic values are close to 1 further supporting the assumption of normality. We can continue with a Pearson correlation test. (For this simple correlation test we don't need to standardize)

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

```
    Pearson's product-moment correlation

data:  df$divorce_rate_maine and df$margarine_consumption_per_capita
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
```

```
#visualise both variables on one plot by year

ggplot(df, aes(x = year)) +
  geom_line(aes(y = scale(divorce_rate_maine),
                color = "Divorce Rate in Maine"), linewidth = 1) +
  geom_line(aes(y = scale(margarine_consumption_per_capita),
                color = "Margarine Consumption per Capita"), linewidth = 1) +
  scale_x_continuous(breaks = 2000:2009) +
  scale_color_manual(values = c("Divorce Rate in Maine" = "palevioletred",
                                "Margarine Consumption per Capita" = "goldenrod1")) +
  labs(
    title = "Divorce Rate and Margarine Consumption in Maine (2000-2009)",
    subtitle = "Standardized (z-score) values for comparison",
    x = "Year", y = "Standardized Value (z)"
  ) +
  theme_minimal() +
  theme(legend.title = element_blank(), legend.position = "top")
```



A Pearson correlation coefficient was computed to assess the linear relationship between divorce rate and margarine consumption in Maine. There was a statistically significant positive correlation between the two variables, $r(8) = .99$, $p < .001$, the CI was [.967, .998].

In other words, there seems to be a positive correlation between margarine consumption and divorce rates in Maine, meaning the changes in margarine is consumption are associated with changes in divorce rates.

However, correlation does not mean causation, simply that margarine consumption and divorce rates seem to change together. There could be other factors that influence both variables, or it could be a coincidence.

Part 2

Work with the 'GSSvocab' dataset from the 'car' package, focusing only on the year 1978 (subset that year and exclude missing values). Investigate how vocabulary test scores (vocab) are related to education (educ). Present the relationship in a plot, fit a regression model, and report results in APA format (coefficients, p-values, and R2), followed by a short practical interpretation.

Extend the analysis by making a new model, where you include whether a person is native- born (nativeBorn) as a predictor. Again, visualise and model the relationship. Consider whether education and native-born status interact in predicting vocabulary, and if so, fit an interaction model. For each model, report results in APA format and explain the practical meaning.

Finally, compare the models and discuss which one performs best based on p-values and R2, as well as the real-world interpretability of the result.

```
df2 <- as.data.frame(GSSvocab)

#subsetting
df2 <- df2 %>%
  filter(year == 1978) %>%
  drop_na()
```

```r
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
```

Dataset

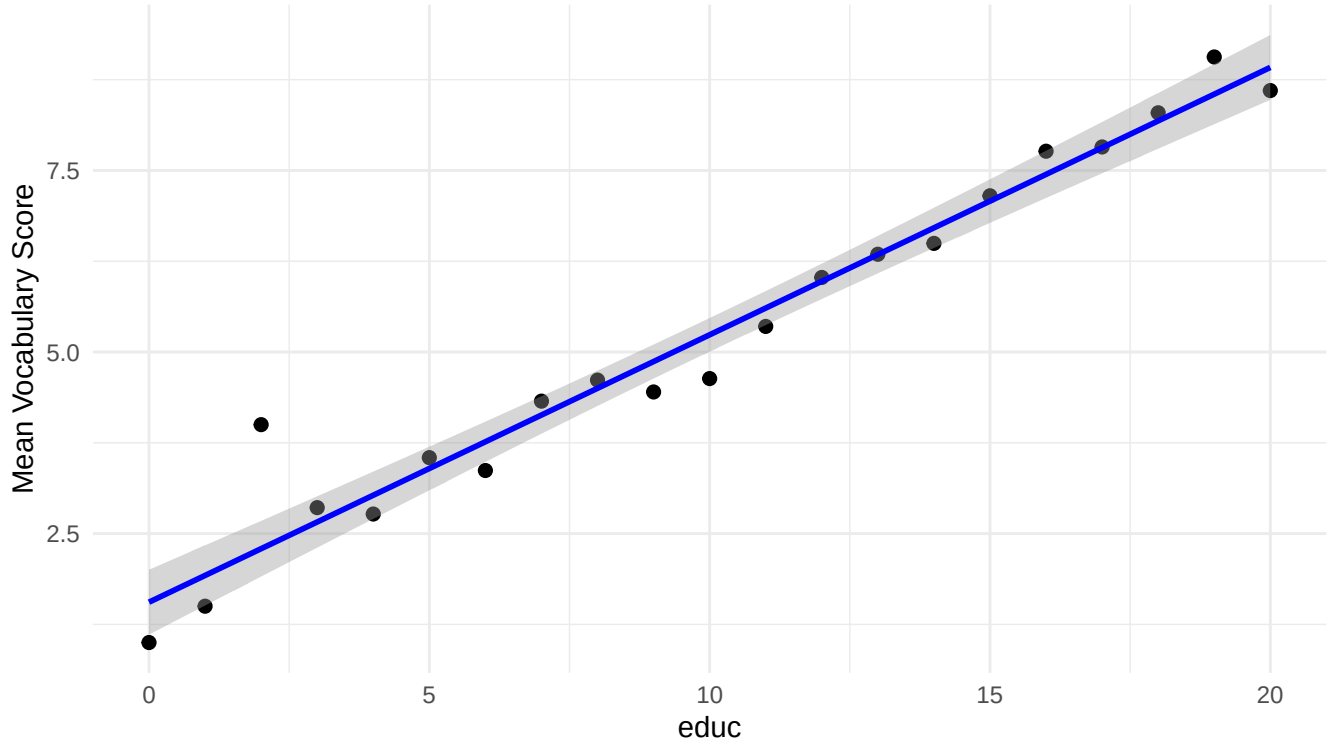
Columns: year: year of the survey, gender: gender, nativeBorn: whether the respondent is native-born, ageGroup: age group of the respondent, educGroup: education group of the respondent, vocab: vocabulary test score, age: age of respondent, educ: years of education.

This subset of the dataset contains 1477 observations from the year 1978.

```r
#checking types of variables
str(df2)
```

```
'data.frame':   1477 obs. of  8 variables:
 $ year      : Factor w/ 20 levels "1978","1982",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ gender    : Factor w/ 2 levels "female","male": 1 1 2 1 1 2 2 2 1 2 ...
 $ nativeBorn: Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ ageGroup  : Factor w/ 5 levels "18-29","30-39",..: 4 5 2 4 3 1 1 4 3 1 ...
 $ educGroup : Factor w/ 5 levels "<12 yrs","12 yrs",..: 2 1 1 2 2 2 2 2 4 2 ...
 $ vocab     : num  10 6 4 9 6 6 4 7 8 3 ...
 $ age       : num  52 74 35 50 41 19 19 59 49 21 ...
 $ educ      : num  12 9 10 12 12 12 12 12 16 12 ...
```

```r
df2 %>%
  group_by(educ) %>%
  summarise(mean_vocab = mean(vocab)) %>%
  ggplot(aes(x = educ, y = mean_vocab)) +
  geom_point(size = 2) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  theme_minimal() +
  labs(y = "Mean Vocabulary Score")
```




```
model1 <- lm(vocab ~ educ, data = df2)
summary(model1)
```

```
Call:
lm(formula = vocab ~ educ, data = df2)

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
```

A simple linear regression showed that education significantly predicted vocabulary scores, $F(1, 1475) = 597.50$, $p < .001$, explaining 28.8% of the variance ($R^2 = .29$). Each additional year of education was associated with an increase of 0.39 points in vocabulary score ($b = 0.39$, SE $= 0.02$, $t = 24.44$, $p < .001$).

```
model2 <- lm(vocab ~ educ + nativeBorn, data = df2)
summary(model2)
```

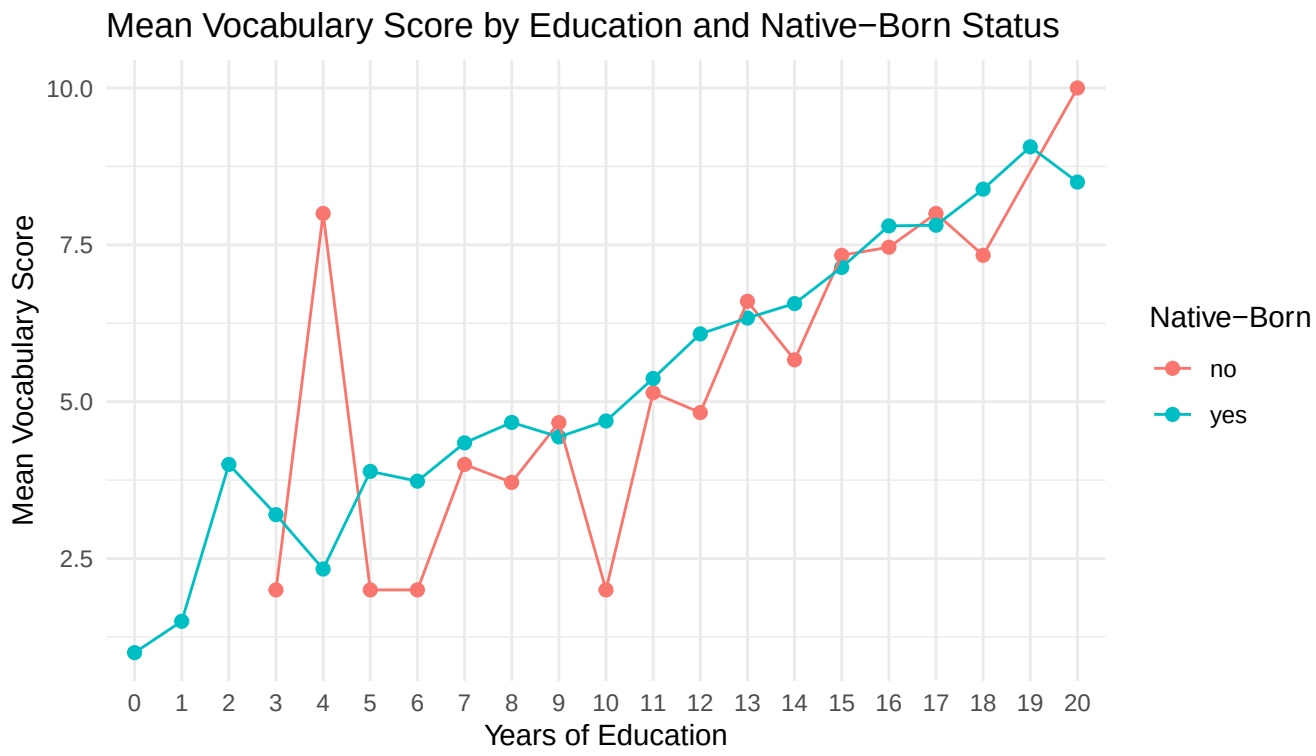
```
Call:
```

```
lm(formula = vocab ~ educ + nativeBorn, data = df2)

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
```

A multiple linear regression was conducted to test whether education and being native-born predict vocabulary scores. The overall model was significant, $F(2, 1474) = 305.60$, $p < .001$, explaining 29.3% of the variance in vocabulary scores ($R^2 = .29$, adjusted $R^2 = .29$). Education was a significant predictor ($b = 0.39$, SE $= 0.02$, $t = 24.50$, $p < .001$). This means each additional year of education was associated with a 0.39-point increase in vocabulary score. Native-born respondents scored on average 0.65 points higher than non-native respondents ($b = 0.65$, SE $= 0.21$, $t = 3.16$, $p = .002$), holding education constant.

To conclude, vocabulary scores seem to increase with education, and even after accounting for education, native-born individuals tend to score slightly higher than foreign-born individuals.

```r
df2 %>%
  group_by(educ, nativeBorn) %>%
  summarise(mean_vocab = mean(vocab), .groups = "drop") %>%
  ggplot(aes(x = factor(educ), y = mean_vocab, color = nativeBorn, group = nativeBorn)) +
  geom_point(size = 2) +
  geom_line() +
  theme_minimal() +
  labs(
    title = "Mean Vocabulary Score by Education and Native-Born Status",
    x = "Years of Education",
    y = "Mean Vocabulary Score",
    color = "Native-Born"
  )
```




```
model3 <- lm(vocab ~ educ * nativeBorn, data = df2)
summary(model3)
```

```
Call:
lm(formula = vocab ~ educ * nativeBorn, data = df2)

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
```

Last but not least, a linear regression with an interaction between education and native-born status was fitted to predict vocabulary scores. The overall model was significant, $F(3, 1473) = 203.70$, $p < .001$, explaining 29.3% of the variance ($R^2 = .293$, adjusted $R^2 = .292$). Education remained a significant predictor of vocabulary scores ($b = 0.42$, SE = 0.05, $p < .001$). Native-born status alone was not a significant predictor after accounting for the interaction ($b = 0.95$, $p = .186$). Crucially, the interaction term between education and native-born status was not significant ($b = –0.03$, SE = 0.06, $t = –0.44$, $p = .663$), indicating that the

effect of education on vocabulary does not differ between native-born and non-native respondents.

Education seems to predict vocabulary scores for both native-born and non-native individuals. Being native-born does not meaningfully change how education influences vocabulary performance.

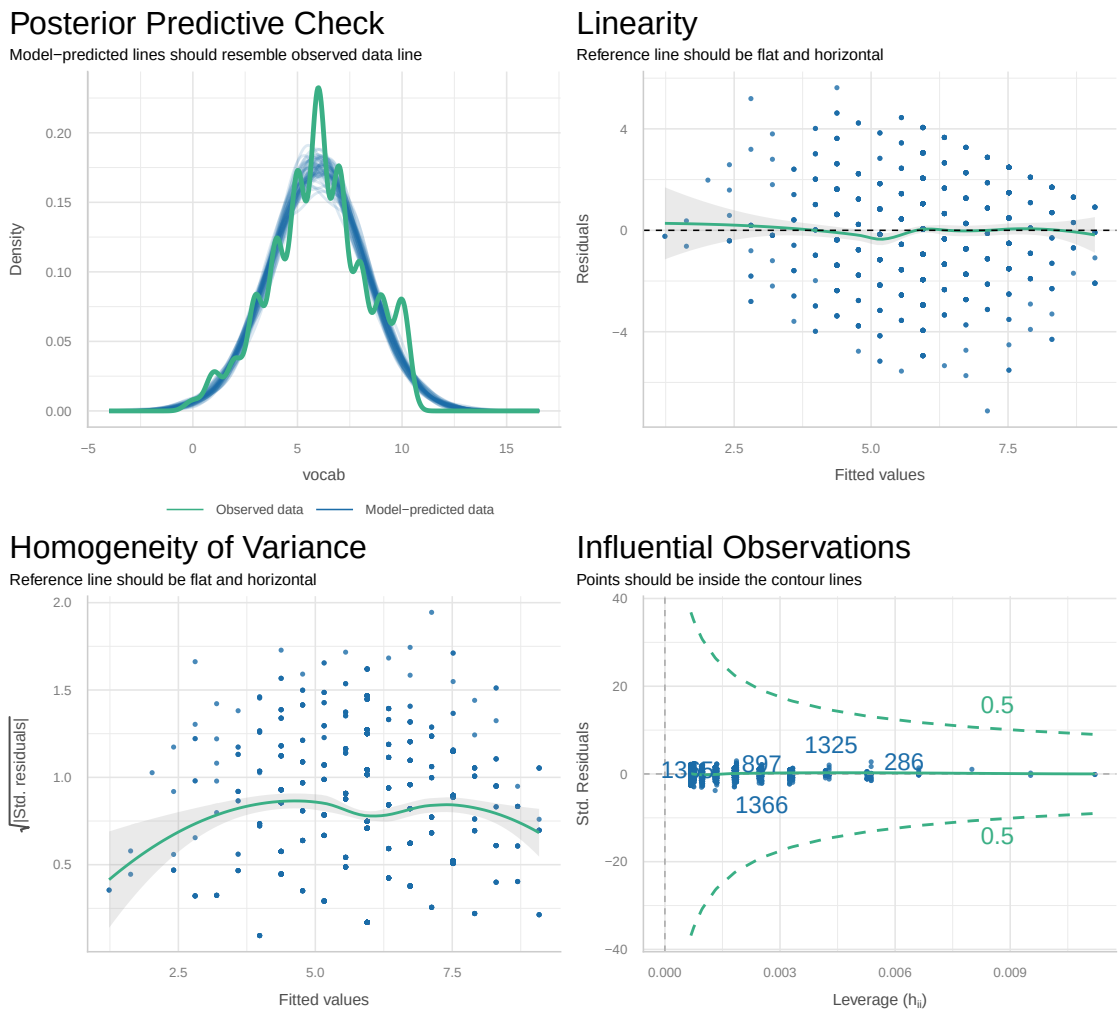
```
anova(model1, model2, model3)
```

```
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
```

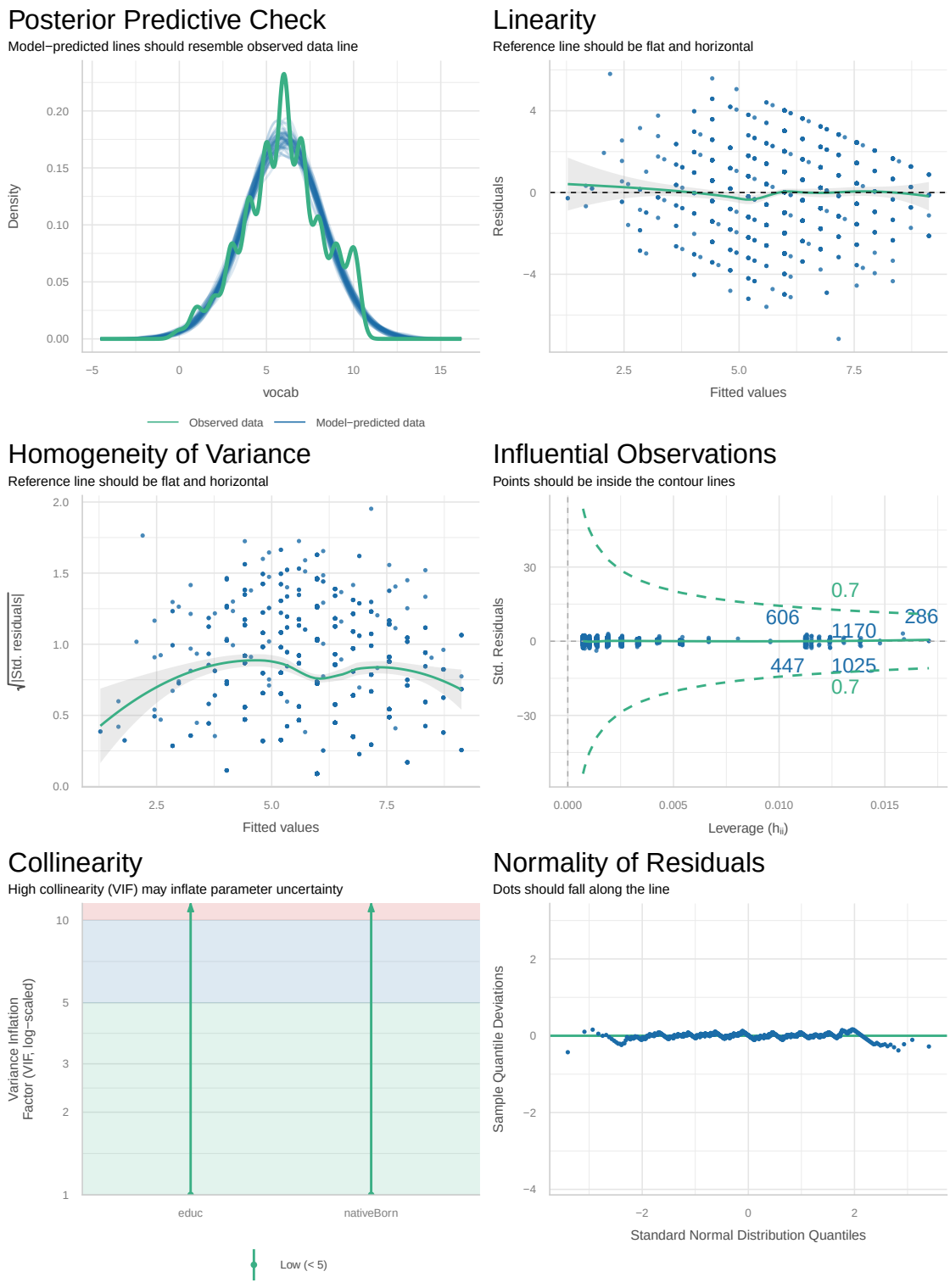
An ANOVA was conducted to compare the three previous models. It shows, that adding native-born status significantly improved model fit over education alone, $F(1, 1474) = 10.01$, $p = .0016$, reducing the residual sum of squares slightly. But including the interaction between education and native-born status did not significantly improve the model, $F(1, 1473) = 0.19$, $p = .663$, with almost no change in the residual sum of squares.

Checking model assumptions

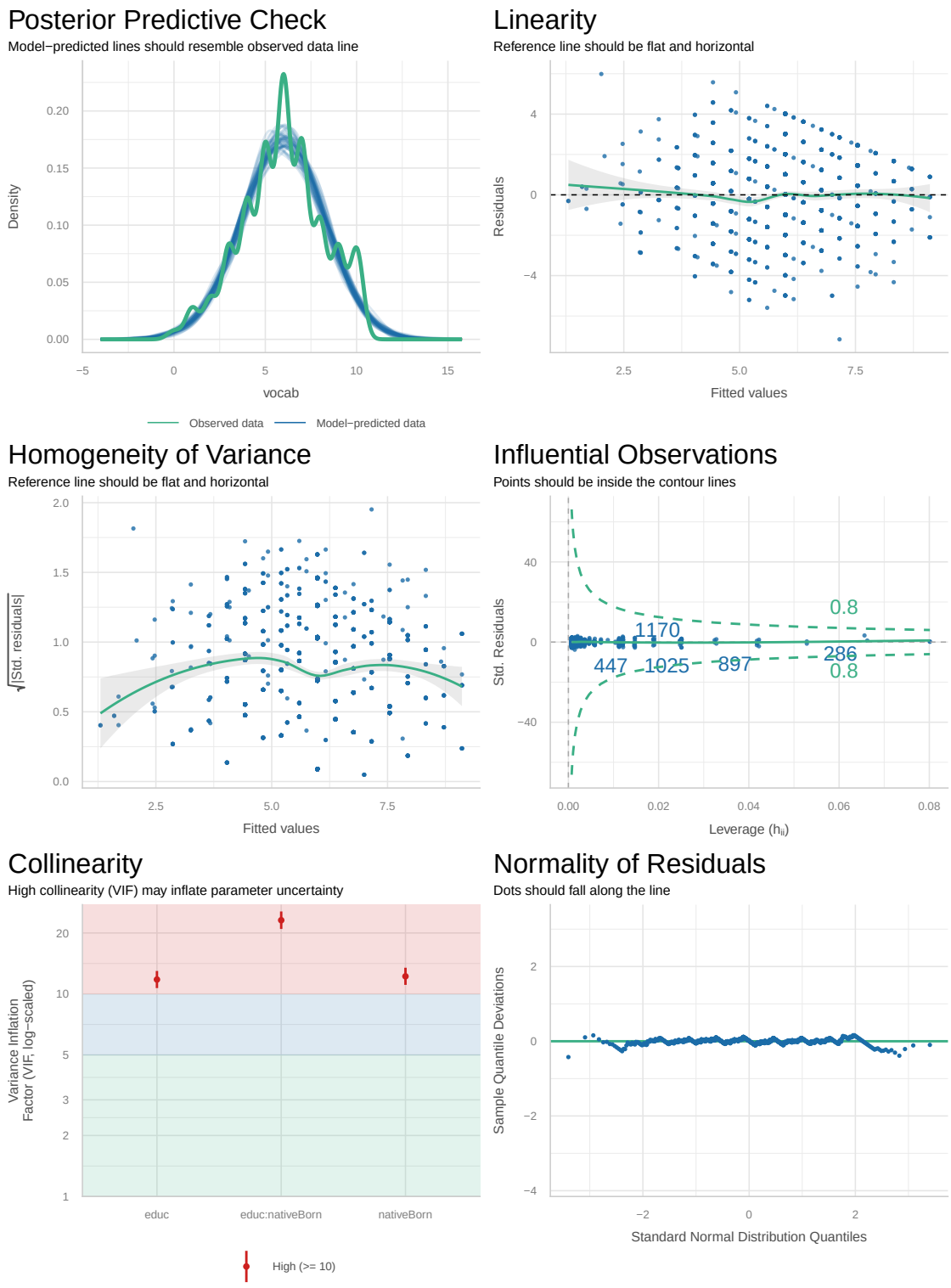
```
check_model(model1, base_size=6, size_dot=0.9, size_line=0.5)
```




```
check_model(model2, base_size=6, size_dot=0.9, size_line=0.5)
```




```
check_model(model3, base_size=6, size_dot=0.9, size_line=0.5)
```



Residuals showed no major violations of model assumptions for model 1, 2 or 3. For the homogeneity of variance plots, shows a slight nonlinear pattern which can be assumed to likely be due to the discrete nature of the variables. Therefore, the assumption of equal variance is considered acceptable.

Short results description From these results, we could say that education can be a predictor of vocabulary scores - with each additional year of schooling linked to higher vocabulary scores,- and being native-born explains a small, but significant additional amount of this variance. However, the effect of education does not differ between native-born and non-native individuals (model 3). Furthermore, many other variables, like social class, income, parental education, cognitive ability, etc could explain part of the relationship between education and vocabulary scores. We also cannot infer causality, since higher education may lead to higher vocabulary, but it's also possible that individuals with higher language ability stay in education longer.