Assignment 02

Réka Forgó

2025-10-15

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)
```

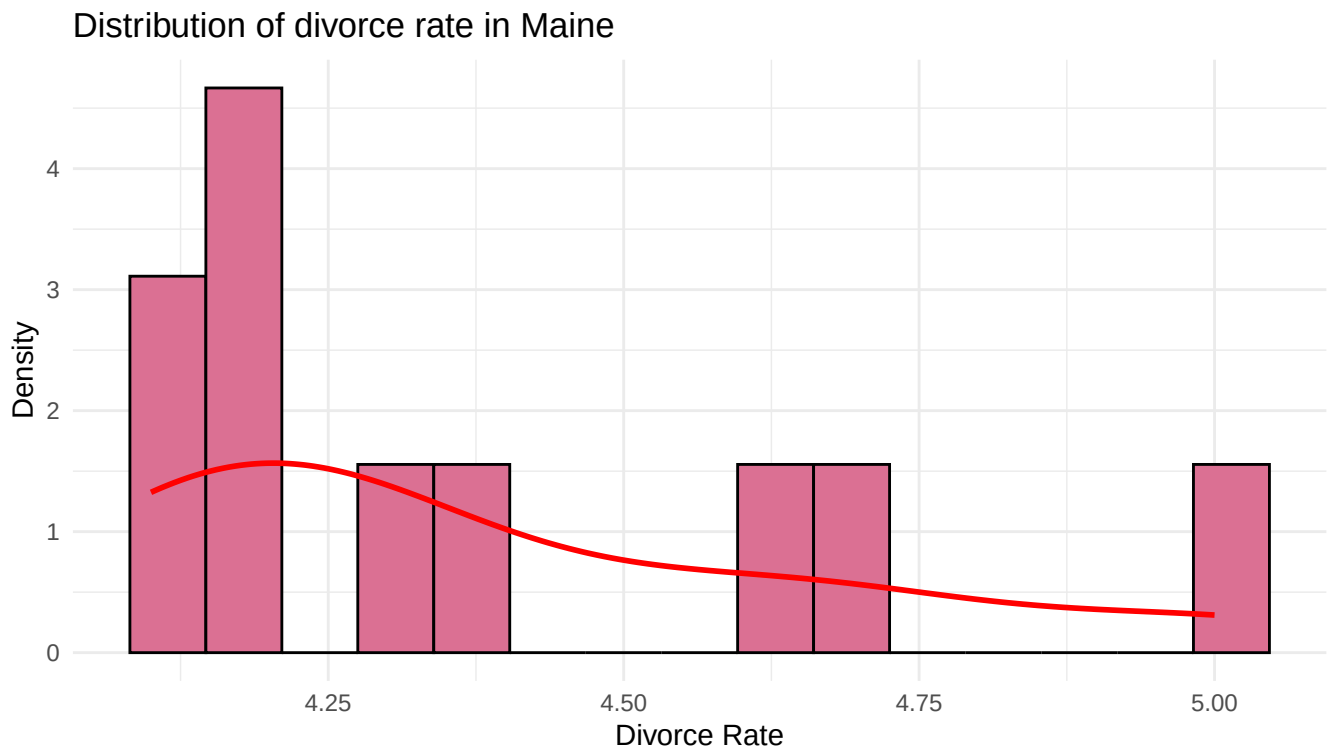
```
df <- as.data.frame(divorce_margarine)
head(df)
```

```
  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
```

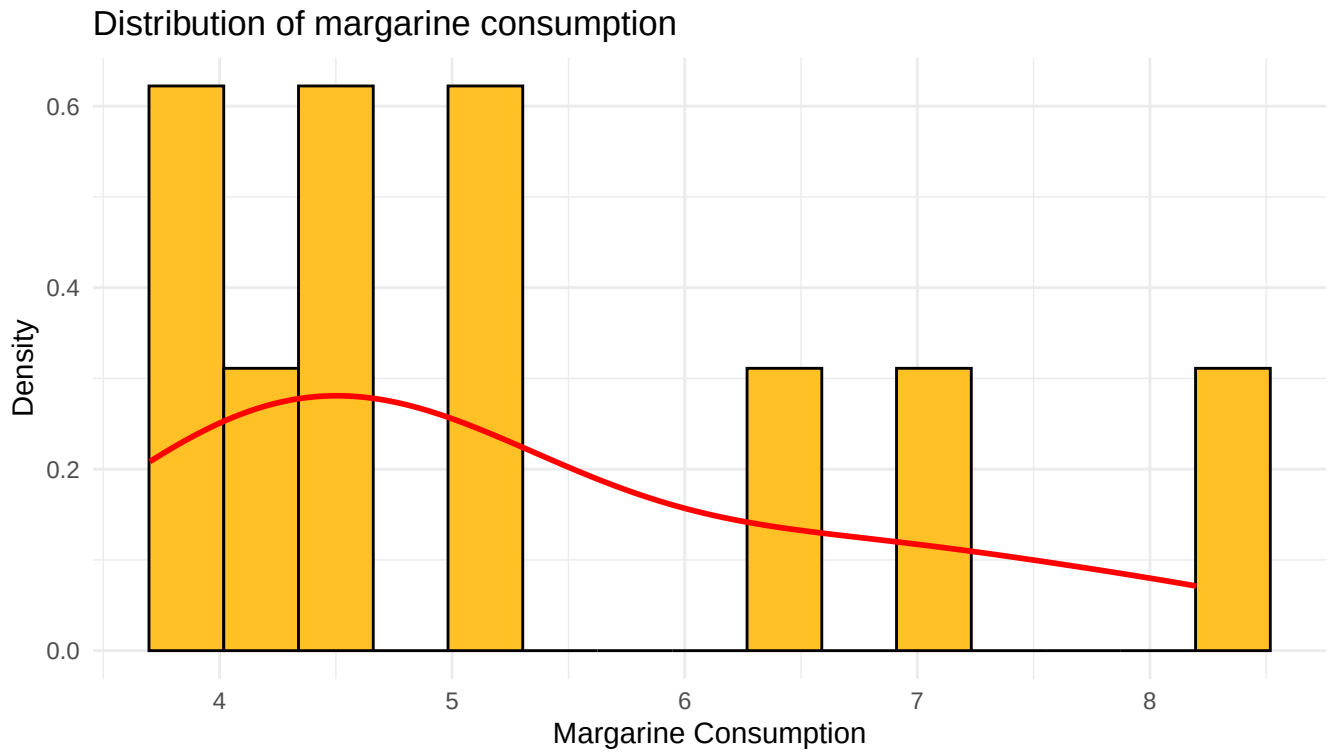
The dataset:

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.

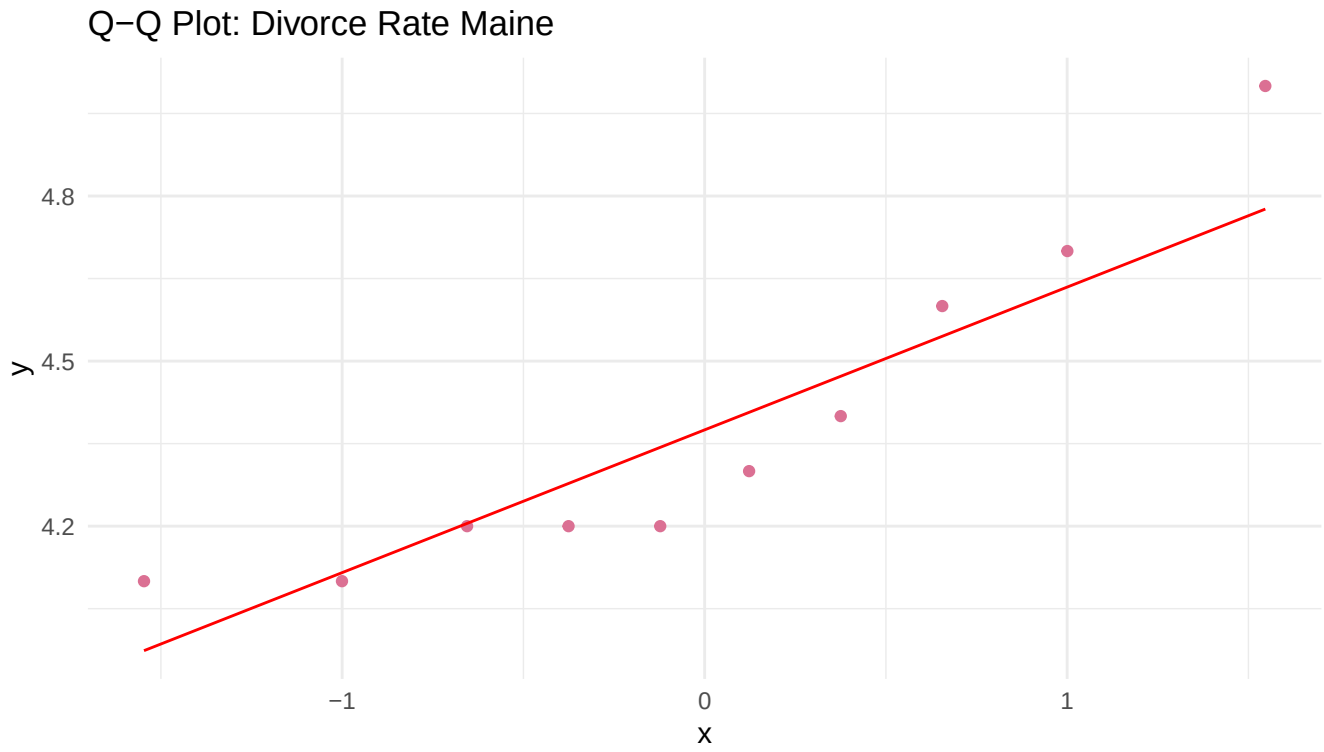
```
#normality checks

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()
```



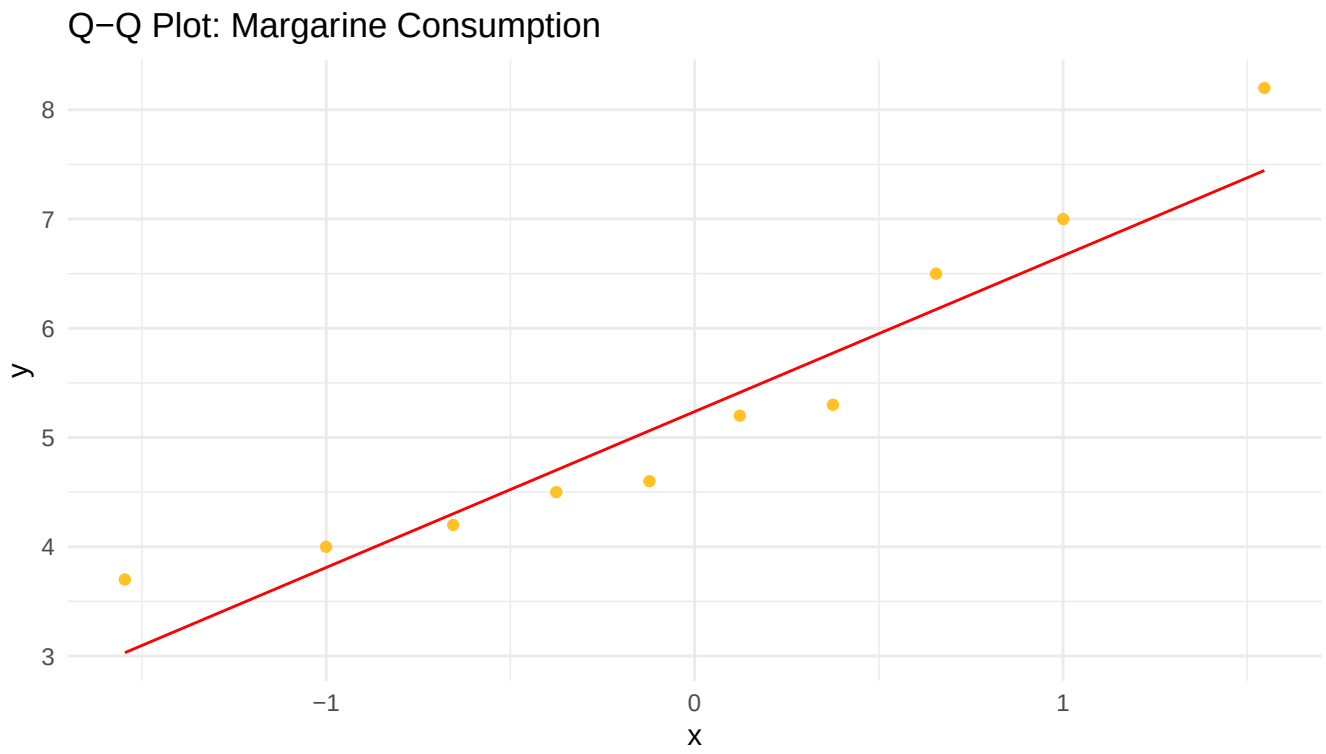

```
ggplot(df, aes(x = margarine_consumption_per_capita)) +
  geom_histogram(aes(y = ..density..), bins = 15,
                 fill = "goldenrod1", color = "black") +
  geom_density(color = "red", linewidth = 1) +
  labs(title = "Distribution of margarine consumption",
       x = "Margarine Consumption",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

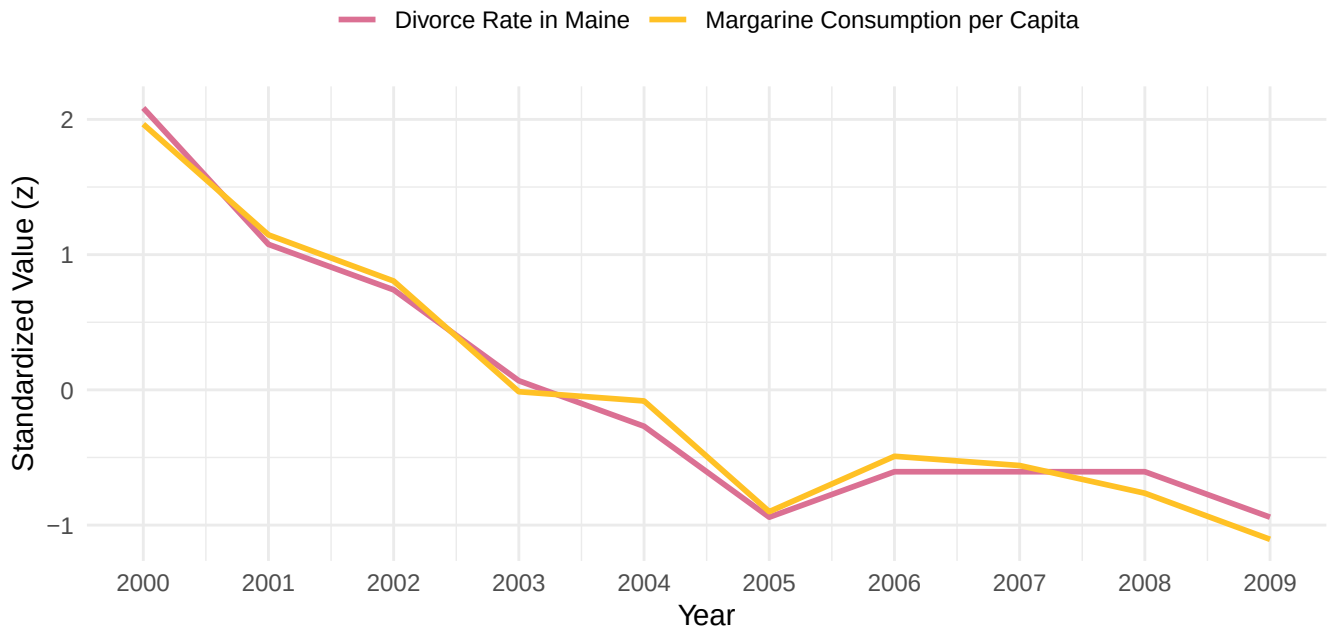
```
coef.var      0.07  0.28
skewness      0.84  0.68
skew.2SE      0.61  0.49
kurtosis     -0.75 -1.03
kurt.2SE     -0.28 -0.39
normtest.W    0.86  0.91
normtest.p    0.08  0.25
```

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 ...
```

We can see that these variables are not continous, and we need to ..