Exercise 8

Tobias Raidl, 11717659

2023-12-12

Contents

```r
library(ggplot2movies)
data(movies)
df = movies[movies$Short == 1 & movies$year >= 2000, ]
```

1

Multiple correlation analysis: Compute the multiple correlation between the variable rating and the matrix consisting of the variables year, length, budget, votes. Delete observations containing missing values. It might be advisable to transform “budget” and “votes”.

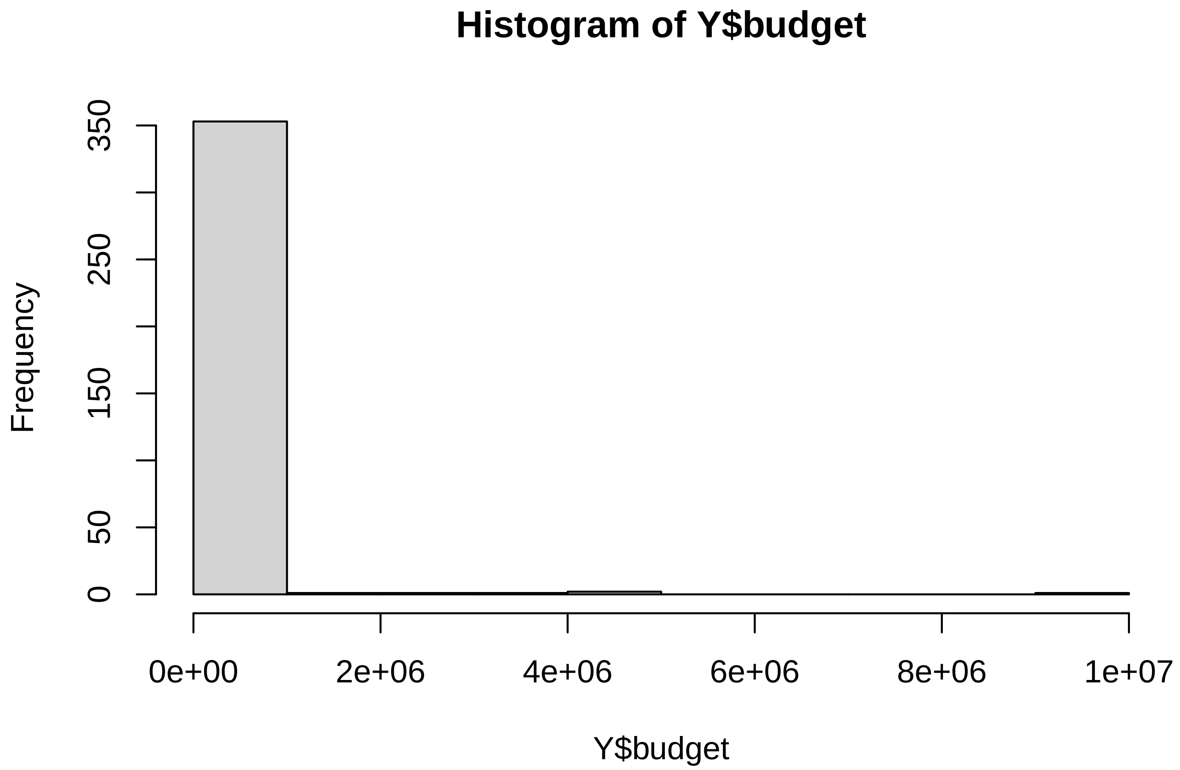
```r
library(dplyr)
```

```
## Warning: Paket 'dplyr' wurde unter R Version 4.3.2 erstellt

##
## Attache Paket: 'dplyr'

## Die folgenden Objekte sind maskiert von 'package:stats':
##
##     filter, lag

## Die folgenden Objekte sind maskiert von 'package:base':
##
##     intersect, setdiff, setequal, union
```

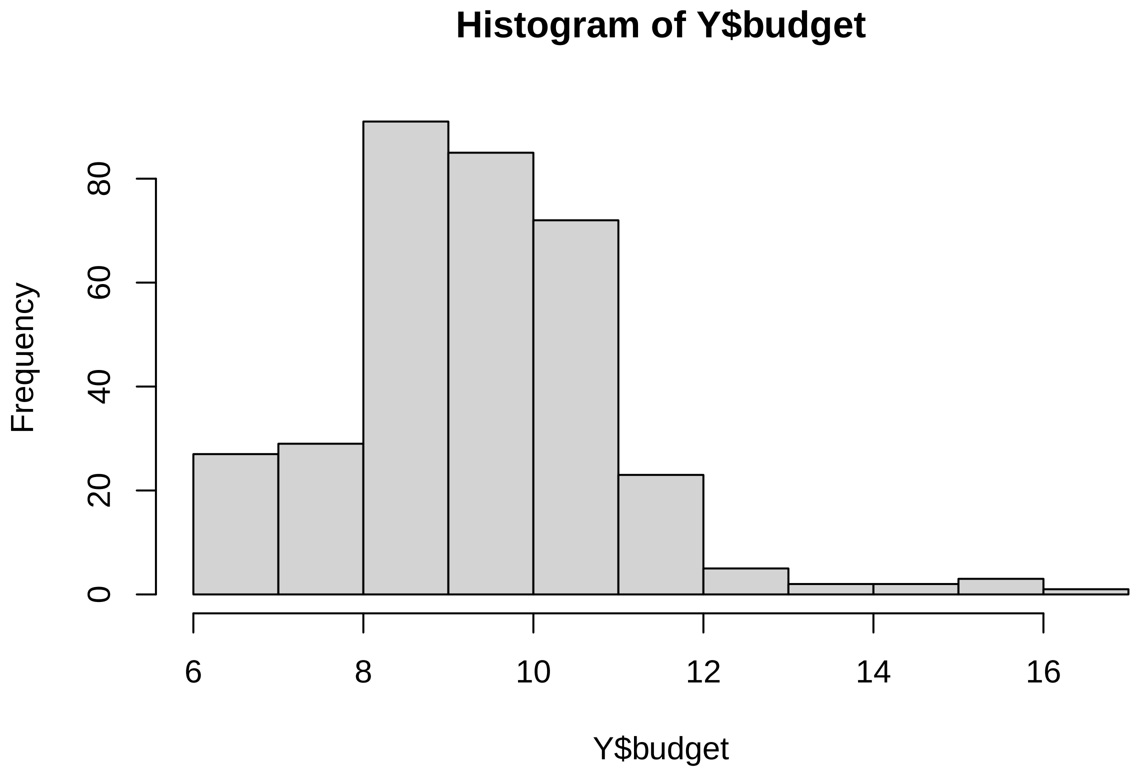
```
df = na.omit(df)
x = df$rating
Y = select(df, c(year, length, budget, votes))
Y
```

```
## # A tibble: 359 x 4
##     year length budget votes
##    <int>  <int>  <int> <int>
##  1  2003     22  32000    11
##  2  2003      9  10000    15
##  3  2005     14   4000    10
##  4  2004     13  12000    11
##  5  2003     13   8000     9
##  6  2003     13   6800     5
##  7  2001     23  13000     8
##  8  2002     27   7000     6
##  9  2003     10   5000     5
## 10  2004     11   5000     6
## # i 349 more rows
```

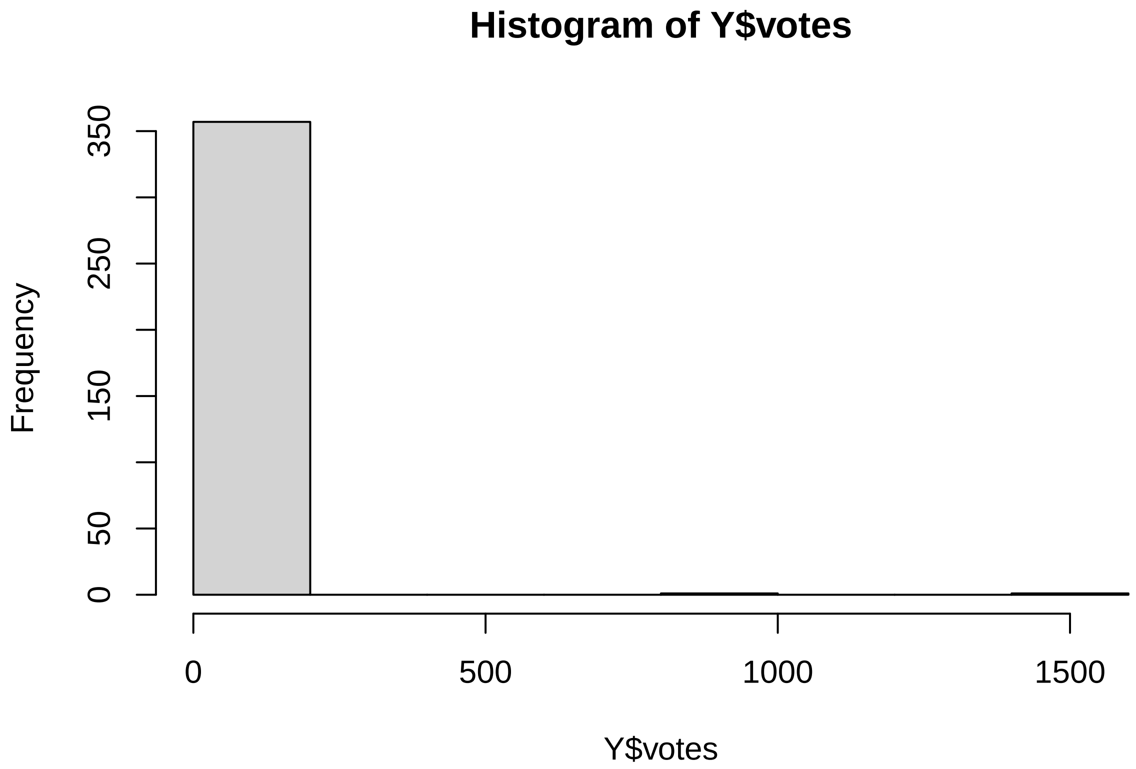
```
hist(Y$budget)
```



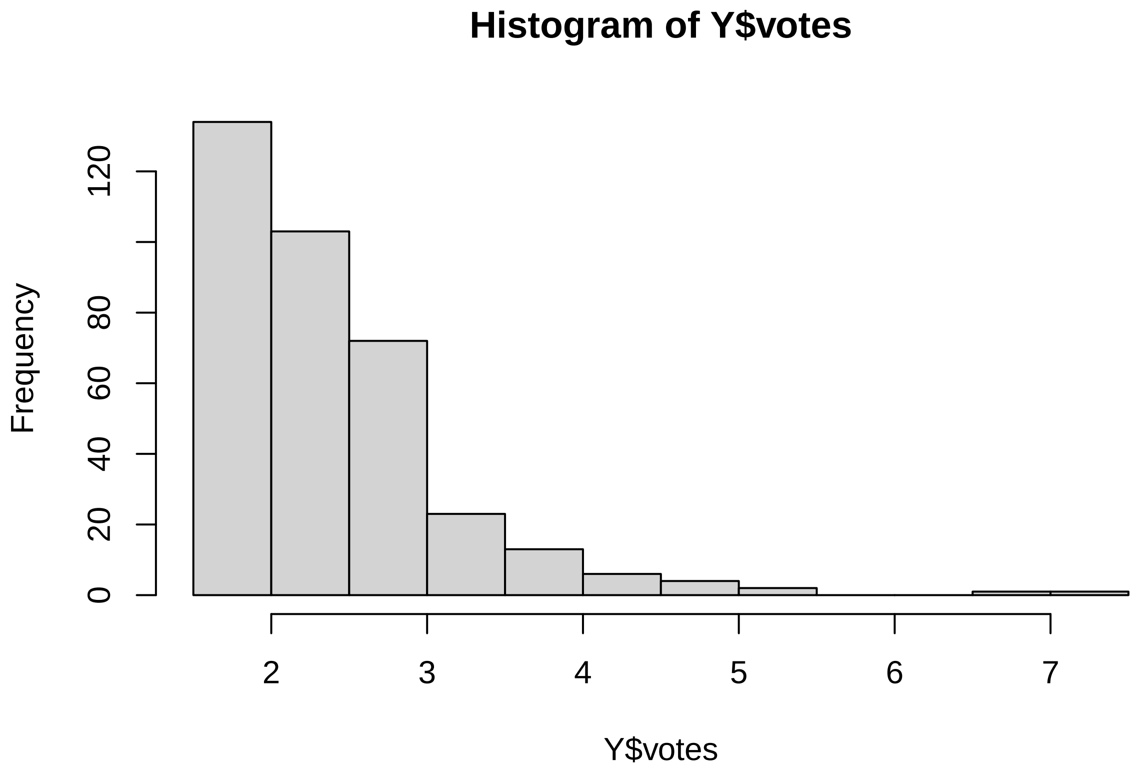
```
Y$budget = log(Y$budget)
hist(Y$budget)
```



```
hist(Y$votes)
```



```
Y$votes = log(Y$votes)
hist(Y$votes)
```



a

Compute the multiple correlation coefficient. How can you interpret the coefficients of the linear predictor function?

```
covmat_xY = cov(cbind(x, Y))
covmat_YY = covmat_xY[-1, -1]
cov_xY = covmat_xY[-1, 1]
str(t(cov_xY))
```

```
##  num [1, 1:4] 0.523 0.122 NaN -0.291
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr [1:4] "year" "length" "budget" "votes"
```

```
str(solve(covmat_YY))
```

```
##  num [1:4, 1:4] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:4] "year" "length" "budget" "votes"
##   ..$ : chr [1:4] "year" "length" "budget" "votes"
```

```
str(cov_xY)
```

```
##  Named num [1:4] 0.523 0.122 NaN -0.291
##  - attr(*, "names")= chr [1:4] "year" "length" "budget" "votes"
```

```
numerator = t(cov_xY) %*% solve(covmat_YY) %*% cov_xY
mult_cor_coef = sqrt(numerator/var(x))
```

b

Is the multiple correlation coefficient significantly different from zero?

c

Use the function CCAgrid from the R package ccaPP – see help. Set the argument method="pearson" and compare the results with those from above. Use method="spearman" and compare with the previous results. What is the methodological difference?

```
library(ccaPP)
```

```
## Warning: Paket 'ccaPP' wurde unter R Version 4.3.2 erstellt

## Lade nötiges Paket: parallel

## Lade nötiges Paket: pcaPP

## Warning: Paket 'pcaPP' wurde unter R Version 4.3.2 erstellt

## Lade nötiges Paket: robustbase

## Warning: Paket 'robustbase' wurde unter R Version 4.3.2 erstellt
```

```
CCAgrid(x, Y, method = "pearson")
```

```
##
## Call:
## CCAgrid(x = x, y = Y, method = "pearson")
##
## Canonical correlations:
## [1] 0.2346074
```

```
CCAgrid(x, Y, method = "spearman")
```

```
##
## Call:
## CCAgrid(x = x, y = Y, method = "spearman")
##
## Canonical correlations:
## [1] 0.3127383
```

The Pearson Correlation Coefficient assesses the linear relationship between variables, while the Spearman Correlation Coefficient evaluates the monotonic relationship. Spearman does not assume normally distributed data.

d

Use the function permTest from the library(ccaPP). This function is performing a permutation test for uncorrelatedness, by permuting the observations of the first input. How and why does this work? What is the outcome? Compare with the result in (b).

```
permTest(x, Y)
```

```
##
## Permutation test for no association
##
## r = 0.299874, p-value = 0.000000
## R = 1000 random permuations
## Alternative hypothesis: true maximum correlation is not equal to 0
```

2

Canonical correlation analysis: Compute the canonical correlation between the matrices consisting of the variables year, length, budget, rating, votes and the variables Action, Animation, Comedy, Drama, Documentary, Romance. Select (transform) the observations according to the instructions at the beginning.

a

Use the function cancor() – see help. Center and scale the data (why?). How strong is the linear relationship? How can you interpret the linear combinations for the X and Y data?

```
X2 = select(df, c(year, length, budget, rating, votes))
Y2 = select(df, c(Action, Animation, Comedy, Drama, Documentary, Romance))
X2_scaled = scale(X2)
Y2_scaled = scale(Y2)
cancor = cancor(X2_scaled, Y2_scaled)
```