STAT 32950: Homework 1

Robert Winter

Table of Contents

1 Exercise 1

Download the data `ladyrun24.dat`. Save the dataset in your working directory. The data are on national track records for women, based on Table 1-9 in Johnson and Wichern. Measurements for 100m, 200, and 400m are in seconds, longer distance records are in minutes. Variable names are not included.

```
ladyrun =
 ↪  read.table("C:/Users/rewin/OneDrive/Documents/STAT_32950/Homework/HW1/ladyrun24.dat")
colnames(ladyrun) = c("Country", "100m", "200m", "400m", "800m", "1500m",
 ↪  "3000m", "Marathon")
```

Compute the following (rounded to 2 decimal places) for the dataset.

1.1 Part (a)

Sample means of the variables. Is there any variable for which the mean is not meaningful (same judgment for the following questions)?

There is not a meaningful way to take the "mean" of country names, so we do not consider this variable. The sample mean national records for each distance in the dataset are summarized below.

```
colMeans(ladyrun %>% subset(select = -Country)) %>% round(2)
```

```
    100m     200m     400m     800m    1500m    3000m Marathon
   11.31    23.07    51.82     2.02     4.19     9.07   153.31
```

1.2 Part (b)

Sample covariance matrix and correlation matrix. Just the R command, no need to print the output.

As above, it is not sensible to compute covariances or correlations between country names and national records, so we again focus on the records themselves.

The sample covariance matrix may be generated as follows:

```
ladyrun %>% subset(select = -Country) %>% cov(method = "pearson") %>%
 ↪  round(2)
```

The sample correlation matrix (using Pearson's r) may be generated as follows:

```
ladyrun %>% subset(select = -Country) %>% cor(method = "pearson") %>%
 ↪  round(2)
```

1.3 Part (c)

Sample correlation matrix using Kendall's τ. Just the R command, no need to print the output.

As above, it is not sensible to compute correlations between country names and national records, so we again focus on the records themselves.

The sample correlation matrix using Kendall's τ may be generated as follows:

```
ladyrun %>% subset(select = -Country) %>% cor(method = "kendall") %>%
 ↪  round(2)
```

1.4 Part (d)

Sample correlation matrix using Spearman's ρ. Just the R command, no need to print the output.

As above, it is not sensible to compute correlations between country names and national records, so we again focus on the records themselves.

The sample correlation matrix using Spearman's ρ may be generated as follows:

```
ladyrun %>% subset(select = -Country) %>% cor(method = "spearman") %>%
 ↪  round(2)
```

1.5 Part (e)

All three types of correlation matrix (Pearson, Kendall, Spearman) on the logarithm of the data. Again, just the R command, no need to print the output. Are the results using log-transformed data the same as in (b), (c), and (d)? Why?

```
# Pearson
ladyrun %>% subset(select = -Country) %>% log() %>% cor(method = "pearson")
 ↪  %>% round(2)

# Kendall
ladyrun %>% subset(select = -Country) %>% log() %>% cor(method = "kendall")
 ↪  %>% round(2)

# Spearman
ladyrun %>% subset(select = -Country) %>% log() %>% cor(method =
 ↪  "spearman") %>% round(2)
```

1. The Pearson correlation coefficients *change* from those in Part (b) when we log-transform the data. This is not surprising: since Pearson's correlation coefficient captures the linear relationship between two variables, and the log transformation is nonlinear, the values of each pairwise correlation will change. Mathematically,

$$\frac{\sum_{i=1}^{7}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{7}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{7}(y_i-\bar{y})^2}} \neq \frac{\sum_{i=1}^{7}(\log(x_i)-\overline{\log(x)})(\log(y_i)-\overline{\log(y)})}{\sqrt{\sum_{i=1}^{7}(\log(x_i)-\overline{\log(x)})^2}\sqrt{\sum_{i=1}^{7}(\log(y_i)-\overline{\log(y)})^2}}$$

2. The Kendall correlation coefficients *do not* change when we log-transform the data. Recall that Kendall's τ is calculated based on the number of concordant and discordant pairs of observations $\{(x_i, y_i), (x_j, y_j)\}$, where a pair is concordant if $(x_i - x_j)(y_i - y_j) >$

0 and discordant if $(x_i - x_j)(y_i - y_j) < 0$. But since the log function is monotonic, we have that $x_i - x_j \gtreqless 0 \Leftrightarrow x_i \gtreqless x_j \Leftrightarrow \log(x_i) \gtreqless \log(x_j) \Leftrightarrow \log(x_i) - \log(x_j) \gtreqless 0$. Similarly, $y_i - y_j \gtreqless 0 \Leftrightarrow \log(y_i) - \log(y_j) \gtreqless 0$. As such, $(x_i - x_j)(y_i - y_j) \gtreqless 0 \Leftrightarrow (\log(x_i) - \log(x_j))(\log(y_i) - \log(y_j)) \gtreqless 0$. That is, log-transforming the data preserves concordance: if a pair of observations were concordant or discordant before the data is log-transformed, they will still be concordant or discordant (respectively) after the log-transformation. The numbers of concordant pairs n_c and discordant pairs n_d do not change, so neither does Kendall's τ.

3. The Spearman correlation coefficients *do not* change when we log-transform the data, either. Recall that Spearman's ρ applies the Pearson correlation formula to the ranks of each observation. Since the log function is monotonic, log-transforming the data preserves the rank of each observation: $x_i \gtreqless x_j \Leftrightarrow \log(x_i) \gtreqless \log(x_j)$. Since the ranks of the log-transformed data are the same as the "raw" data, Spearman's ρ does not change.

1.6 Part (f)

Now for the sample correlation matrix R (of all meaningful variables), obtain the eigenvalues (to 2 decimal places) and the eigenvectors (no need for output).

The seven eigenvalues of R are listed below:

```r
eigen(ladyrun %>% subset(select = -Country) %>% cor(method =
 ↪  "pearson"))$values %>% round(2)
```

```
[1] 5.70 0.74 0.29 0.11 0.09 0.05 0.02
```

Eigenvectors corresponding to these eigenvalues may be generated as follows:

```r
eigen(ladyrun %>% subset(select = -Country) %>% cor(method =
 ↪  "pearson"))$vectors %>% round(2)
```

(i) What is the sum of all eigenvalues? Compare it to the dimensions of the variables.

As shown above, the sum of the eigenvalues of R is $\sum_{i=1}^{7} \lambda_i = 7$, which is precisely the number of quantitative variables (i.e., excluding country name) in the dataset.

```r
eigen(ladyrun %>% subset(select = -Country) %>% cor(method =
 ↪  "pearson"))$values %>% sum()
```

```
[1] 7
```

(ii) Give the dimension(s) of each eigenvector.

Each eigenvector has nonzero coordinates in $\mathbb{R}^7$, so each eigenvector has dimension 7.