Homework 6 (Optional)

Insert your name here

Table of contents

[Link to the Github repository](#)

> ❗ Due: Wed, May 3, 2023 @ 11:59pm
>
> Please read the instructions carefully before submitting your assignment.
>
> 1. This assignment requires you to only upload a PDF file on Canvas
> 2. Don't collapse any code cells before submitting.
> 3. Remember to make sure all your code output is rendered properly before uploading your submission.
>
> Please add your name to the author information in the frontmatter before submitting your assignment

In this assignment, we will perform various tasks involving principal component analysis (PCA), principal component regression, and dimensionality reduction.

We will need the following packages:

```r
packages <- c(
  "tibble",
  "dplyr",
  "readr",
  "tidyr",
  "purrr",
  "broom",
```

```
    "magrittr",
    "corrplot",
    "car"
)
# renv::install(packages)
sapply(packages, require, character.only=T)
```

Question 1

> 💡 70 points
>
> Principal component anlaysis and variable selection

1.1 (5 points)

The `data` folder contains a `spending.csv` dataset which is an illustrative sample of monthly spending data for a group of 5000 people across a variety of categories. The response variable, `income`, is their monthly income, and objective is to predict the `income` for a an individual based on their spending patterns.

Read the data file as a tibble in R. Preprocess the data such that:

1. the variables are of the right data type, e.g., categorical variables are encoded as factors
2. all column names to lower case for consistency
3. Any observations with missing values are dropped

```
path <- "data/spending.csv"

df <- ... # Insert your code here
```

1.2 (5 points)

Visualize the correlation between the variables using the `corrplot()` function. What do you observe? What does this mean for the model?

```
df_x  %>% ... # Insert your code here
```

1.3 (5 points)

Run a linear regression model to predict the `income` variable using the remaining predictors. Interpret the coefficients and summarize your results.

```
... # Insert your code here
```

1.3 (5 points)

Diagnose the model using the `vif()` function. What do you observe? What does this mean for the model?

```
... # Insert your code here
```

1.4 (5 points)

Perform PCA using the `princomp` function in R. Print the summary of the PCA object.

```
pca <- ... # Insert your code here
... # Insert your code here
```

1.5 (5 points)

Make a screeplot of the proportion of variance explained by each principal component. How many principal components would you choose to keep? Why?

```
... # Insert your code here
```

1.6 (5 points)

By setting any factor loadings below 0.2 to 0, summarize the factor loadings for the principal components that you chose to keep.

```
clean_loadings <- ... # Insert your code here
```

Visualize the factor loadings.

```
... # Insert your code here
```

1.7 (15 points)

Based on the factor loadings, what do you think the principal components represent?

Provide an interpreation for each principal component you chose to keep.

1.8 (10 points)

Create a new data frame with the original response variable `income` and the principal components you chose to keep. Call this data frame `df_pca`.

```
... # Insert your code here
```

Fit a regression model to predict the `income` variable using the principal components you chose to keep. Interpret the coefficients and summarize your results.

```
... # Insert your code here
```

Compare the results of the regression model in 1.3 and 1.9. What do you observe? What does this mean for the model?

```
... # Insert your code here
```

1.10 (10 points)

Based on your interpretation of the principal components from Question 1.7, provide an interpretation of the regression model in Question 1.9.

> **ⓘ Session Information**
>
> Print your `R` session information using the following command
>
> ```
> sessionInfo()
> ```
>
> ```
> R version 4.2.2 (2022-10-31)
> Platform: x86_64-conda-linux-gnu (64-bit)
> Running under: Storage
>
> Matrix products: default
> BLAS/LAPACK: /gpfs/group/hvt5139/default/tools/mambaforge/lib/libopenblasp-r0.3.21.so
>
> locale:
>  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
>  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
>  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
>  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
>  [9] LC_ADDRESS=C               LC_TELEPHONE=C
> [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
>
> attached base packages:
> [1] stats     graphics  grDevices datasets  utils     methods   base
>
> other attached packages:
>  [1] car_3.1-1        carData_3.0-5   tibble_3.1.8     luz_0.3.1
>  [5] torch_0.9.0      e1071_1.7-13    rpart.plot_3.1.1 rpart_4.1.19
>  [9] caret_6.0-94     lattice_0.20-45 ggplot2_3.4.1    corrplot_0.92
> [13] magrittr_2.0.3   broom_1.0.4     purrr_1.0.1      tidyr_1.3.0
> [17] readr_2.1.4      dplyr_1.1.0
>
> loaded via a namespace (and not attached):
>  [1] nlme_3.1-162     fs_1.6.1         lubridate_1.9.2
>  [4] bit64_4.0.5      progress_1.2.2   tools_4.2.2
>  [7] backports_1.4.1  utf8_1.2.3       R6_2.5.1
> [10] colorspace_2.1-0 nnet_7.3-18      withr_2.5.0
> [13] tidyselect_1.2.0 prettyunits_1.1.1 processx_3.8.0
> [16] bit_4.0.5        compiler_4.2.2   cli_3.6.0
> [19] scales_1.2.1     callr_3.7.3      proxy_0.4-27
> [22] stringr_1.5.0    digest_0.6.31    rmarkdown_2.20
> ```

```
[25] coro_1.0.3            pkgconfig_2.0.3        htmltools_0.5.4
[28] parallelly_1.35.0     fastmap_1.1.0          rlang_1.0.6
[31] generics_0.1.3        jsonlite_1.8.4         ModelMetrics_1.2.2.2
[34] Matrix_1.5-3          Rcpp_1.0.10            munsell_0.5.0
[37] fansi_1.0.4           abind_1.4-5            lifecycle_1.0.3
[40] stringi_1.7.12        pROC_1.18.0            yaml_2.3.7
[43] MASS_7.3-58.3         plyr_1.8.8             recipes_1.0.5
[46] grid_4.2.2            parallel_4.2.2         listenv_0.9.0
[49] crayon_1.5.2          splines_4.2.2          hms_1.1.2
[52] zeallot_0.1.0         knitr_1.42             ps_1.7.2
[55] pillar_1.8.1          future.apply_1.10.0    reshape2_1.4.4
[58] codetools_0.2-19      stats4_4.2.2           glue_1.6.2
[61] evaluate_0.20         data.table_1.14.8      renv_0.16.0-53
[64] vctrs_0.5.2           tzdb_0.3.0             foreach_1.5.2
[67] gtable_0.3.1          future_1.32.0          xfun_0.37
[70] gower_1.0.1           prodlim_2019.11.13     class_7.3-21
[73] survival_3.5-5        timeDate_4022.108      iterators_1.0.14
[76] hardhat_1.2.0         lava_1.7.2.1           timechange_0.2.0
[79] globals_0.16.2        ellipsis_0.3.2         ipred_0.9-14
```