Preoperative Atelectasis

Part 3: Assessment of Independent Variables

Javier Mancilla Galindo

2023-12-01

Table of contents

Setup

Packages used

```
if (!require("pacman", quietly = TRUE)) {
  install.packages("pacman")
}


pacman::p_load(
  tidyverse, # Used for basic data handling and visualization.
  table1, #Used to add lables to variables.
  mgcv, #Used to model non-linear relationships with a general additive model.
  ggmosaic, #Used to create mosaic plots.
  car, #Used to visualize distribution of continuous variables (stacked Q-Q plots).
  dagitty, #Used in conjunction with https://www.dagitty.net/ to create
          #directed acyclic graph to inform statistical modelling.
  report #Used to cite packages used in this session.
)
```

Session and package dependencies

```
R version 4.3.2 (2023-10-31 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 11 x64 (build 22621)

Matrix products: default


locale:
[1] LC_COLLATE=Spanish_Mexico.utf8  LC_CTYPE=Spanish_Mexico.utf8
[3] LC_MONETARY=Spanish_Mexico.utf8 LC_NUMERIC=C
[5] LC_TIME=Spanish_Mexico.utf8

time zone: Europe/Berlin
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices datasets  utils     methods   base

other attached packages:
 [1] report_0.5.7    dagitty_0.3-1   car_3.1-2       carData_3.0-5
 [5] ggmosaic_0.3.3  mgcv_1.9-0      nlme_3.1-164    table1_1.4.3
 [9] lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
[13] purrr_1.0.2     readr_2.1.4     tidyr_1.3.0     tibble_3.2.1
```

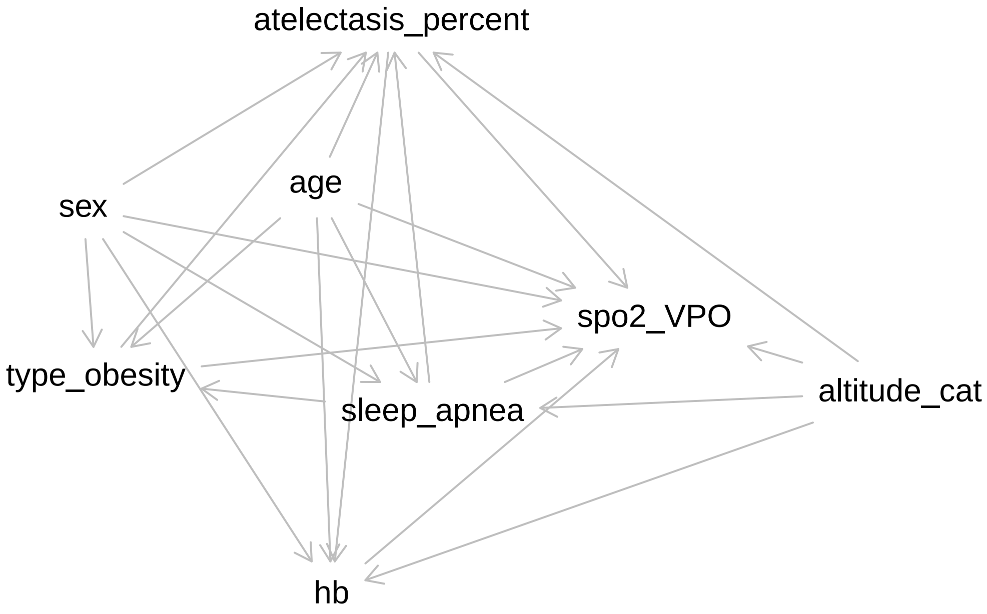
```
[17] ggplot2_3.4.4   tidyverse_2.0.0 pacman_0.5.1
```

Assessment of independent variables

The selection of variables that will be assessed is according to the following directed acyclic graph which will be used again before statistical modelling, to assess conditional independencies.

DAG

DAG generated in the DAGitty website and sourced from the accompanying script ***DAG.R***



Other variables that are potential confounders are not shown in this DAG since they were addressed by design in this study as follows:

- Current COVID-19: Exclusion criteria were applied to **n=2** patients with CO-RADS 3 and **n=2** with CO-RADS 4. Only participants with low probability of COVID-19 (CO-RADS 1 and 2) were included in this study.
- Prior COVID-19: This was an exclusion criterion (**n=3**).
- Bronchiectasis: This was an exclusion criterion (n=0).

Description of independent variables

Age

Summary:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  20.00   32.75   40.00   40.26   48.25   65.00 
```

The mean age was 40.3 (SD: 9.87).

Sex

Frequencies:

```
sex
  Man Woman 
   22   214 
```

Percentage:

```
sex
  Man Woman 
  9.3  90.7 
```

Most patients in the sample were woman (n=22, 9.3%).

Body mass index (BMI)

Summary:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  30.00   34.63   40.30   41.37   46.02   77.31 
```

Frequencies:

```
type_obesity
Class 1 Obesity Class 2 Obesity Class 3 Obesity 
             62              53             121 
```

Percentage:

```
type_obesity
Class 1 Obesity Class 2 Obesity Class 3 Obesity 
           26.3            22.5            51.3 
```

Distribution of BMI was assessed earlier. It is right-skewed due to extreme values (verified outliers). The WHO classification of BMI for obesity class will be used to complement descriptions and for potential use later during statistical modelling.

> The median BMI was 40.295 (IQR: 34.63- 46.02). The distribution of BMI was right-skewed due to extreme BMI values (range: 30- 77.31). Most patients were in the class 3 obesity category (n=121, 51.3%), followed by class 1 (n=62, 26.3%) and 2 (n=53, 22.5%). a

SpO2

Summary:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     88      93      96      95      97      99 
```

Distribution of SpO2 during the pre-anesthetic is left-skewed due to some participants exhibiting decreased SpO2. I will categorize according to clinical categories to assess the proportion of patients with decreased SpO2:

Proportion of patients with decreased SpO2

Frequencies:

```
spo2_cat
     90 90 to 94      >94 
      15       75      146 
```

Percentage:

```
spo2_cat
     90 90 to 94      >94 
     6.4     31.8     61.9 
```

> The median SpO2 during the pre-anethetic assessment was 96 (IQR: 93-97) %, with a minimum value of 88%. Of these, n=146 (61.9%) had normal SpO2 (above 94%), whereas n=75 (31.8%) had a value in the 90-94% range, and n=15 (6.4%) had 90%.

Obstructive sleep apnea

Frequencies:

```
sleep_apnea
 No Yes
218  18
```

Percentage:

```
sleep_apnea
  No  Yes
92.4  7.6
```

Patients with a diagnosis of OSA were 7.6% (n=18) of the sample.

Altitude

Summary:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   31.0   519.0   519.0   652.7   806.0  1861.0
```

Distribution of altitude was assessed earlier. Distribution is very unclear due to very widespread datapoints. Thus, I will create a new variable categorizing values according to the study by Crocker ME, et al.

Frequencies:

```
altitude_cat
     Low altitude Moderate altitude
              205                31
```

Percentage:

```
altitude_cat
     Low altitude Moderate altitude
             86.9              13.1
```

Hemoglobin

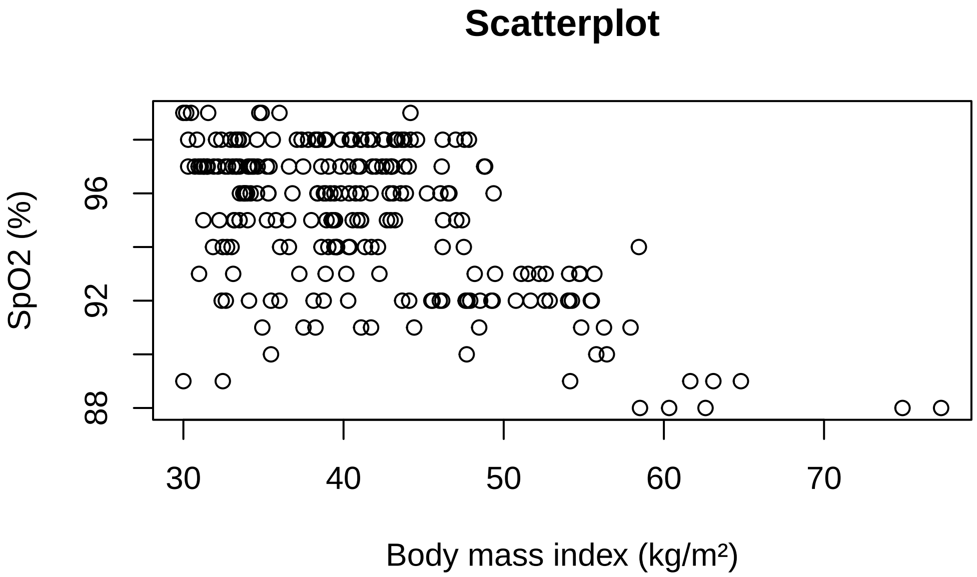
Summary:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   9.90   13.90   14.50   14.54   15.20   18.50       2 
```

Distribution of hemoglobin was assessed and follows a normal distribution. Two participants don't have a hemoglobin value.

Relationships between independent variables

BMI and SpO2



Relationship does not seem to be linear (also, variables were not normally distributed, with outliers), but suggests a negative correlation. Will assess if a smooth BMI term explains SpO2 better, and if so, what is the best number of knots to model this relationship:

Models evaluated with the accompanying sourced script ***nonlinear_BMI_SpO2.R***

All non-linear models are significantly better than linear. Thus, using a smooth term for BMI is better than modelling a linear relationship.

Best AIC:

```r
list(AIC_k2,AIC_k4,AIC_k6,AIC_k8,AIC_k12)
```

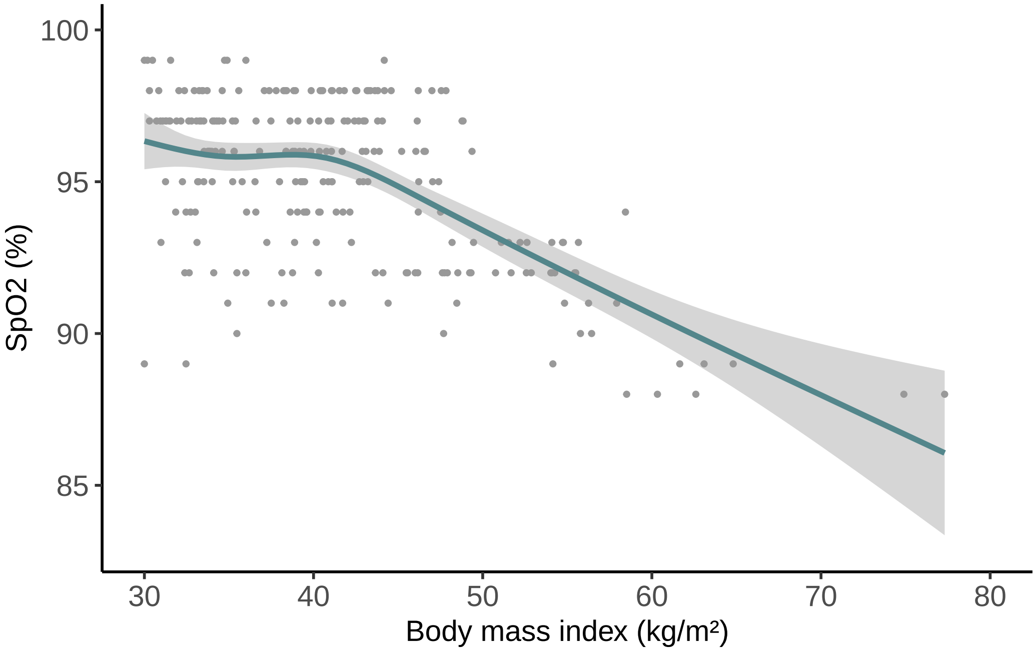
```
[[1]]
[1] 1048.14

[[2]]
[1] 1040.448

[[3]]
[1] 1036.959

[[4]]
[1] 1036.83

[[5]]
[1] 1037.165
```

Regarding AIC, the models with k>6 are not better at explaining the variance. Thus, I will with k=5 since the best model is expected to be anywhere between k=4 and k=6:

```r
list(AIC_k4,AIC_k5,AIC_k6)
```

```
[[1]]
[1] 1040.448

[[2]]
[1] 1037.475

[[3]]
[1] 1036.959
```

Model with k=5 still offers and advantage compared to k=4 (drop in AIC). No other improvements in k-index or visual representation are achieved with higher k. Thus, will use k=5 to model.



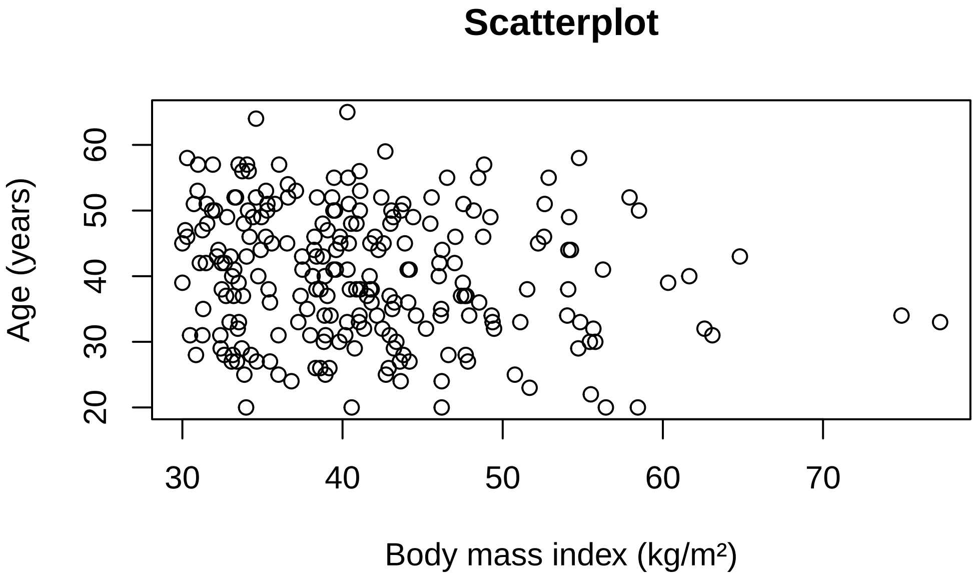
Negative non-monotonic relationship since SpO2 decreases, but then seems to increase slightly again at BMI 40, followed by a marked decrease as BMI decreases at values higher than ~42.

Spearman's correlation coefficient shouldn't be used due to relationship not being monotonically decreasing. However, I will calculate it just to have a rough idea (but will not report this in the paper).

```
    Spearman's rank correlation rho

data:  spo2_VPO and BMI
S = 3105183, p-value = 2.28e-11
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.417458
```

> BMI exhibited a negative non-linear monotonic relationship with SpO2 (**Figure 1B**, rho= -0.417, p<0.001).

BMI and age



Datapoints scattered. Relationship monotonic and probably linear, but there are influential true outliers with extreme BMI. Will assess with Spearman correlation analysis due to extreme BMI values.

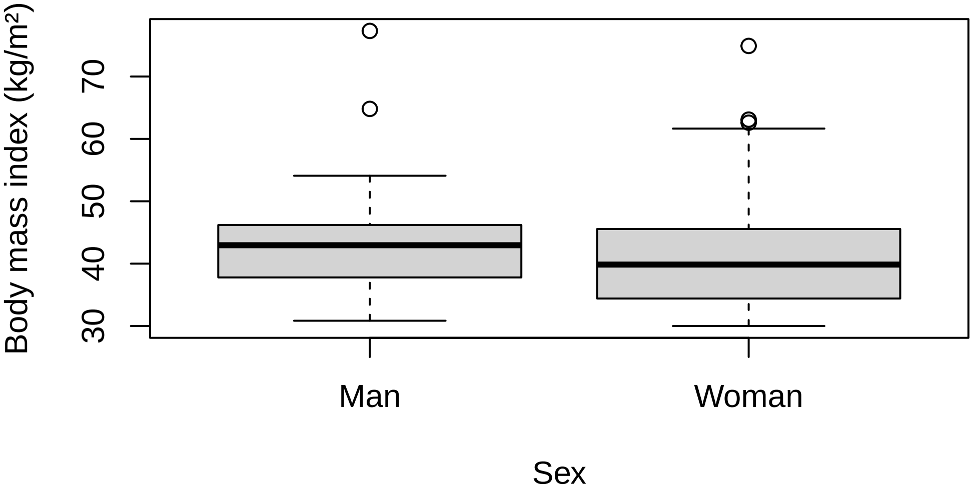
```
	Spearman's rank correlation rho

data:  age and BMI
S = 2530759, p-value = 0.017
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho 
-0.1552445 
```

Age had a weak negative correlation with BMI (rho= -0.155, p=0.017).

BMI and sex

Median BMI:

```
# A tibble: 2 x 7
  sex       n median    Q1    Q3   min   max
  <fct> <int>  <dbl> <dbl> <dbl> <dbl> <dbl>
1 Man      22   43.0  37.9  46.2  30.8  77.3
2 Woman   214   39.8  34.5  45.5  30    74.9
```

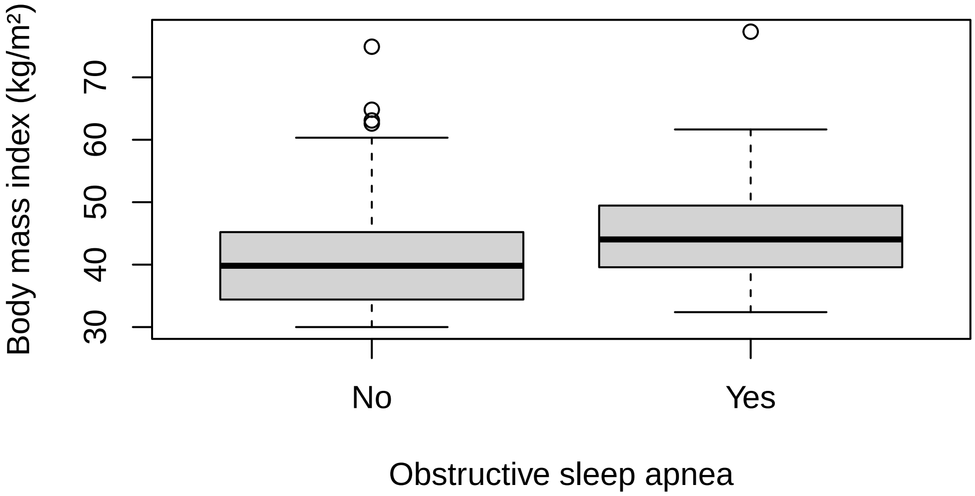


Distribution not normal and influential outliers. Will assess non-parametrically.

```
    Wilcoxon rank sum test with continuity correction

data:  BMI by sex
W = 2789.5, p-value = 0.1537
alternative hypothesis: true location shift is not equal to 0
```

> The median BMI was not different between men (39.9, IQR: 34.5-45.5) and women (43, IQR: 37.9-46.2) (p=0.154).

BMI and sleep apnea

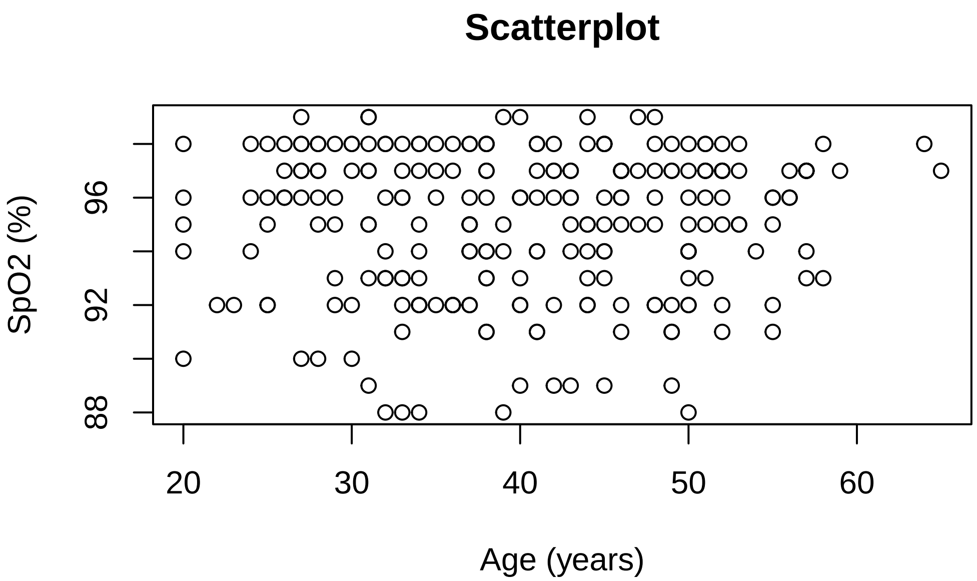


Distribution not normal and influential outliers. Will assess non-parametrically.

```
# A tibble: 2 x 7
  sleep_apnea     n median    Q1    Q3   min   max
  <fct>       <int>  <dbl> <dbl> <dbl> <dbl> <dbl>
1 No            218   39.8  34.5  45.1  30    74.9
2 Yes            18   44.0  40.1  49.2  32.4  77.3

    Wilcoxon rank sum test with continuity correction

data:  BMI by sleep_apnea
W = 1274, p-value = 0.01353
alternative hypothesis: true location shift is not equal to 0
```

> The median BMI was significantly higher in participants with sleep apnea (44, IQR: 40.1-49.2) compared to those without OSA (39.8, IQR: 34.5-45.1) (p=0.014).

Age and SpO2

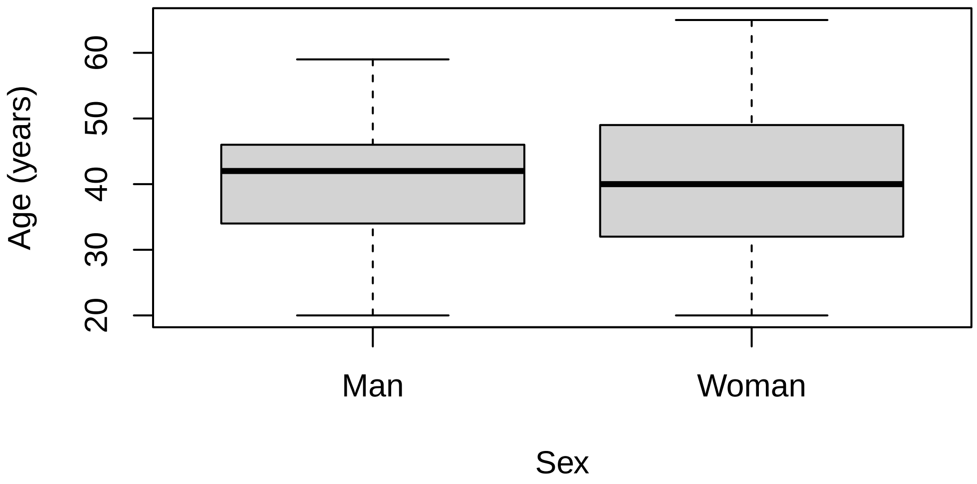


Do not seem to be correlated. Will apply Spearman's correlation test:

```
    Spearman's rank correlation rho

data:  spo2_VPO and age
S = 2143192, p-value = 0.7405
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02167287
```

Age and SpO2 were not correlated (rho= 0.022, p=0.74).

Age and sex



Distribution near-normal, but light tails for women. However, t-test could be robust to deviations from normality and differences in group size. Will assess mean and variance for further testing:

```
# A tibble: 2 x 5
  sex         n age_mean    sd variance
  <fct> <int>    <dbl> <dbl>    <dbl>
1 Man      22     40.6  9.28     86.1
2 Woman   214     40.2  9.94     98.9
```

Variances are similar. However, group sizes differ my 10x. Welch's t-test more suitable:

```

    Welch Two Sample t-test

data:  age by sex
t = 0.19917, df = 26.213, p-value = 0.8437
alternative hypothesis: true difference in means between group Man and group Woman is not equ
95 percent confidence interval:
 -3.882438  4.715913
sample estimates:
  mean in group Man mean in group Woman
           40.63636            40.21963
```

> Mean age was similar bethween men (40.2, sd:9.9) and women (40.6, sd:9.3) (p=0.844).

Age and sleep apnea

Distribution near-normal. Will assess mean and variance for further testing.

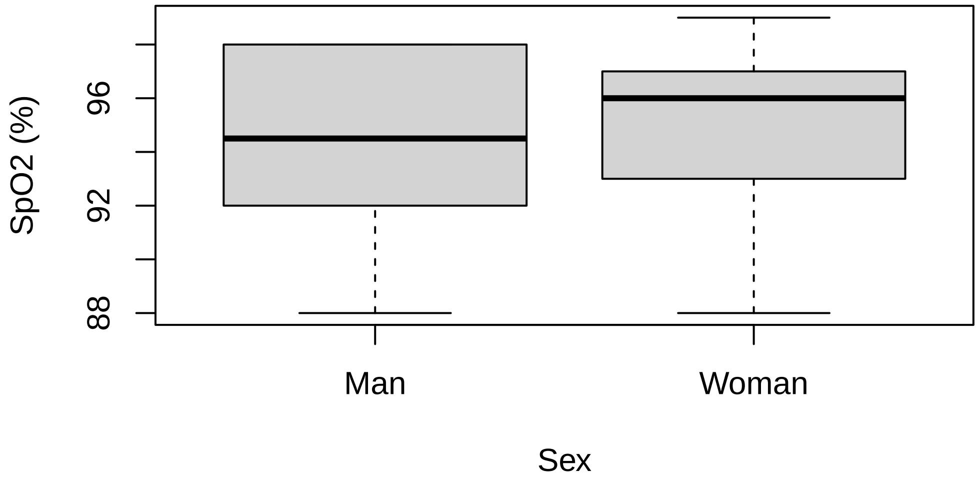
```
# A tibble: 2 x 5
  sleep_apnea     n age_mean    sd variance
  <fct>       <int>    <dbl> <dbl>    <dbl>
1 No            218     40.2 10.0     100.
2 Yes            18     41.4  8.19     67.1
```

Size per group very different, variances do not look similar. Welch's t-test more suitable:

```
    Welch Two Sample t-test

data:  age by sleep_apnea
t = -0.59817, df = 21.42, p-value = 0.556
alternative hypothesis: true difference in means between group No and group Yes is not equal
95 percent confidence interval:
 -5.473186  3.025683
sample estimates:
 mean in group No mean in group Yes
         40.16514          41.38889
```

> Age was not significantly different between participants with OSA (41.4, sd:8.2) and those without (40.2, sd:10) (p=0.556).

SpO2 and sex



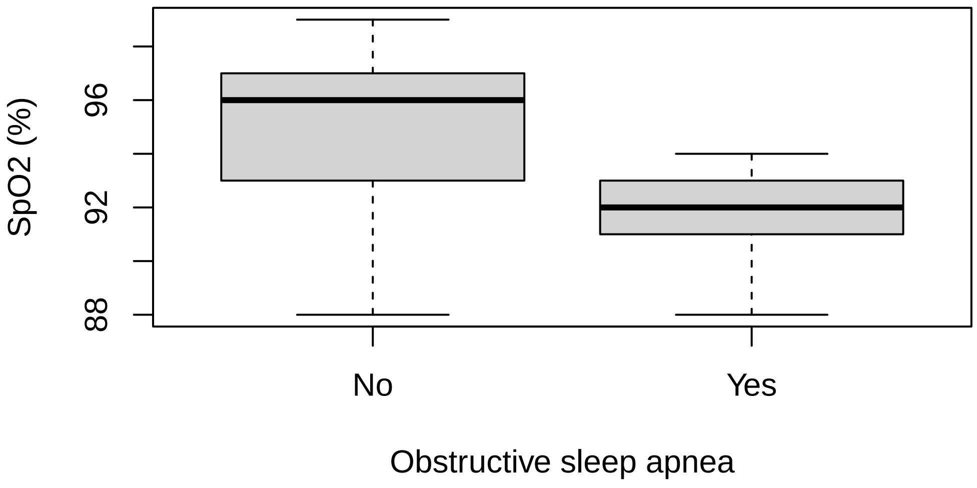
Distribution deviates from normal and small group size for men. Will assess non-parametrically.

```
# A tibble: 2 x 7
  sex       n spo2_median    Q1    Q3   min   max
  <fct> <int>       <dbl> <dbl> <dbl> <int> <int>
1 Man      22        94.5    92  97.8    88    98
2 Woman   214        96      93  97      88    99
```

```
    Wilcoxon rank sum test with continuity correction

data:  spo2_VPO by sex
W = 2106, p-value = 0.413
alternative hypothesis: true location shift is not equal to 0
```

> The median SpO2 was not different between men (96, IQR: 93-97) and women (94.5, IQR: 92-97.8) (p=0.413).

SpO2 and sleep apnea



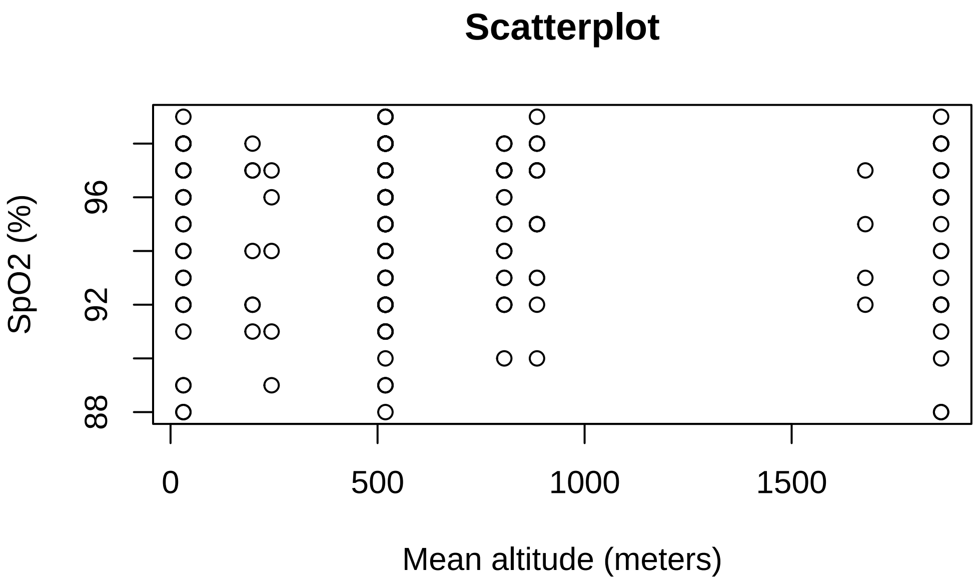
Distribution not normal, and smaller group size for those with sleep apnea. Will assess non-parametrically.

```
# A tibble: 2 x 7
  sleep_apnea     n spo2_median    Q1    Q3   min   max
  <fct>       <int>       <dbl> <dbl> <dbl> <int> <int>
1 No            218          96    93    97    88    99
2 Yes            18          92    91    93    88    94
```

```
    Wilcoxon rank sum test with continuity correction

data:  spo2_VPO by sleep_apnea
W = 3350, p-value = 4.973e-07
alternative hypothesis: true location shift is not equal to 0
```

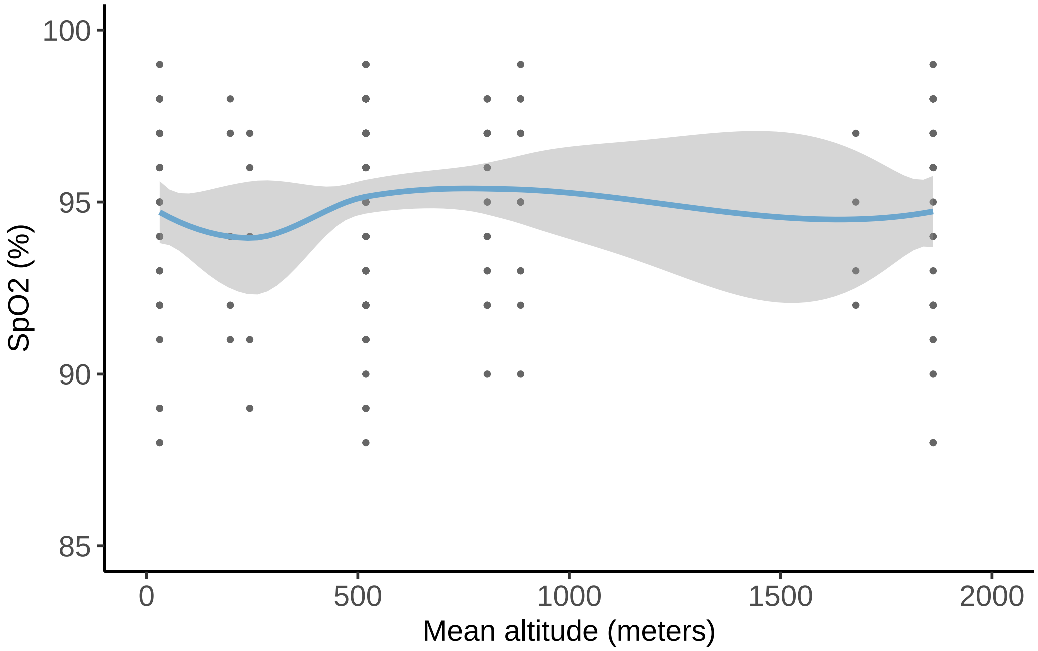
> Patients with sleep apnea had a lower median SpO2 (92, IQR: 91-93) than those without OSA (96, IQR: 93-97) (p<0.001).

SpO2 and altitude



There does not seem to be a pattern.

Would a smooth term be useful to model altitude?



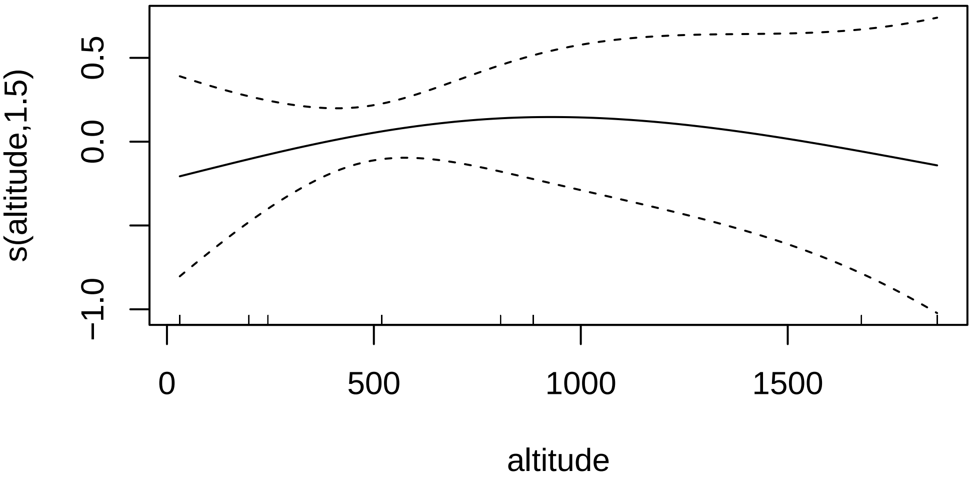
It is likely that a smooth term for SpO2 would be non-informative since there is no clear reasonable pattern in this smooth plot. Additionally, it is well known that any impacts in SpO2 due to altitudes up to 2000 are very limited (i.e 1 to 2 units). go to reference.

I will still check if a smooth term may be better than linear in case that adjustment for this variable is needed.

GAM model with k=4 (this was also checked with varying k from 2 to 10):

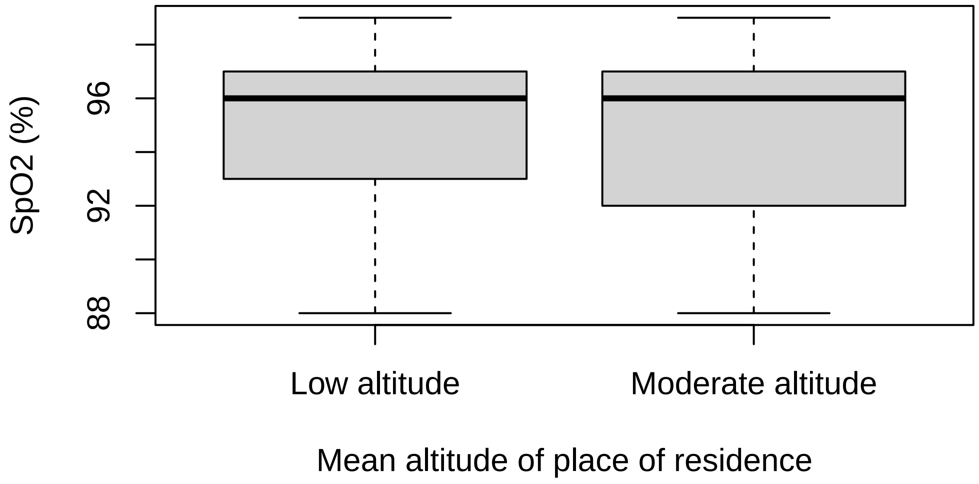
```
Family: gaussian
Link function: identity

Formula:
spo2_VPO ~ s(altitude, k = 4)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   94.996      0.178   533.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
              edf Ref.df     F p-value
s(altitude) 1.505  1.798 0.437   0.631
```

```
R-sq.(adj) =  0.000124   Deviance explained = 0.653%
GCV = 7.5559  Scale est. = 7.4757    n = 236
```



Smooth term is not significantly better than one assuming linearity. Furthermore, the relationship with SpO2 in smooth term does not make any sense (i.e., according to prior reference, SpO2 should decrease at higher altitudes). Thus, it would be very likely that including this term would only explain noise in any case, not the true known causal relationship between SpO2 and altitude.

Lastly, will check the pattern according to altitude categories, which may be a better term to use in models in any case.



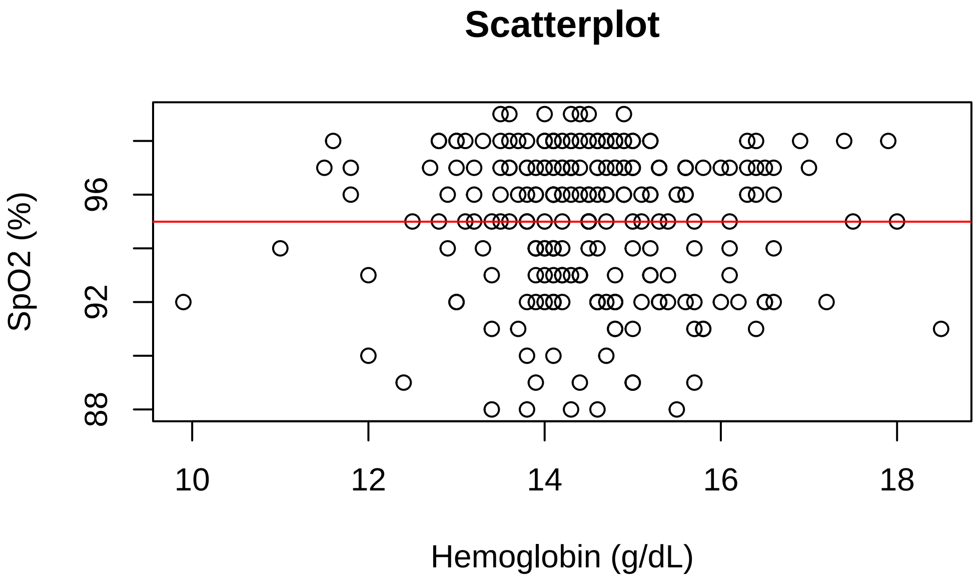
Distribution deviates from normal and small group size for the moderate altitude group. Will assess non-parametrically.

```
# A tibble: 2 x 7
  altitude_cat          n spo2_median    Q1    Q3   min   max
  <fct>             <int>       <int> <dbl> <dbl> <int> <int>
1 Low altitude        205          96    93    97    88    99
2 Moderate altitude    31          96    92    97    88    99
```

```
    Wilcoxon rank sum test with continuity correction

data:  spo2_VPO by altitude_cat
W = 3360, p-value = 0.6043
alternative hypothesis: true location shift is not equal to 0
```

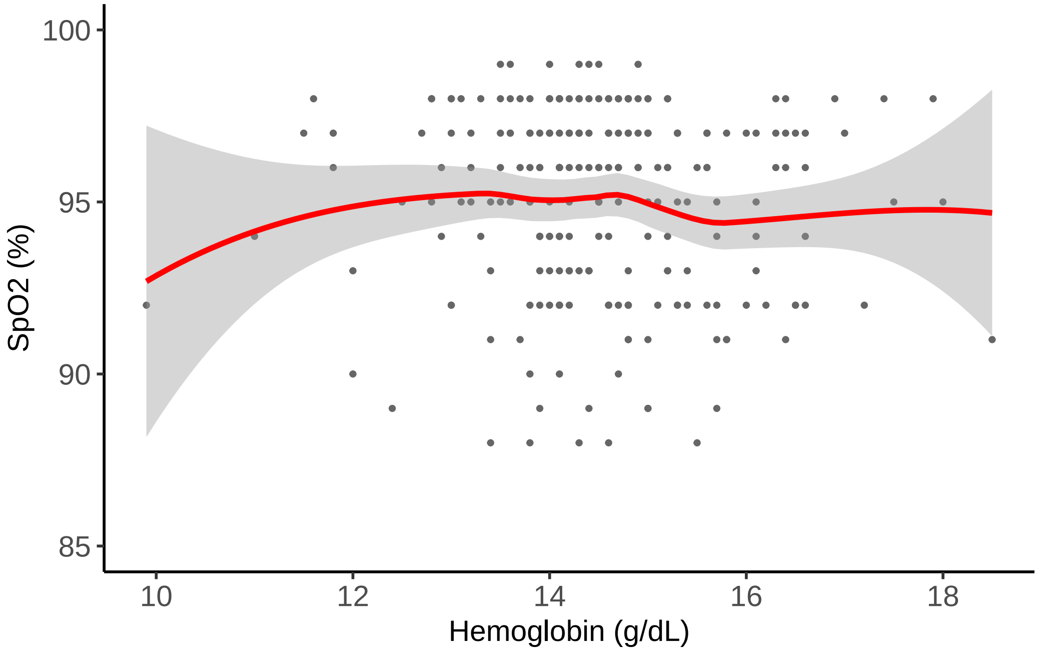
> The median SpO2 was not different between low and moderate altitude categories (p=0.604).

SpO2 and hemoglobin



There does not seem to be a clear pattern.

Would a smooth term be useful to model SpO2?

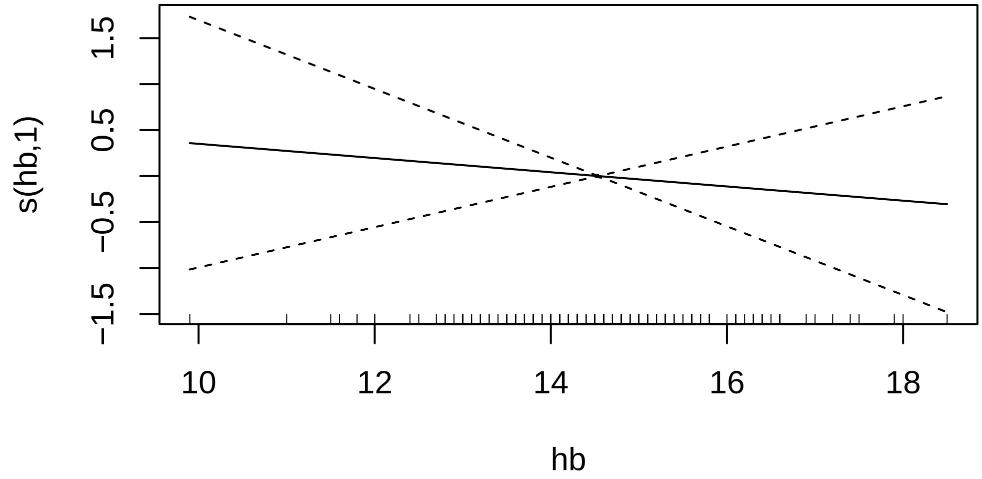


Hemoglobin likely has an effect on SpO2 at lower hemoglobin values, which makes sense with what is observed in the graph. Assuming a linear relationship could lead to incorrect conclusions according to this. Nonetheless, it looks like the apparent non-linear relationship at low Hb values is due to only 2 observations with wide confidence intervals showing that the true slope could go either up, straight or down, so it may also be incorrect to assume a non-linear relationship based only on this plot. I will model to see if there is an optimal smooth term for hemoglobin or if a linear term best fits the data:

GAM model with k=4 (this was also checked with varying k from 2 to 10):

```
Family: gaussian
Link function: identity

Formula:
spo2_VPO ~ s(hb, k = 4)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  94.9829     0.1789   530.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
      edf Ref.df     F p-value
```

```
s(hb)   1      1 0.272   0.603

R-sq.(adj) =  -0.00314   Deviance explained = 0.117%
GCV = 7.5555  Scale est. = 7.4909    n = 234
```



The estimated degrees of freedom (edf) in both cases were 1, plus p=0.6, meaning that a linear term is better fitted to this data than a non-linear term.

```
    Spearman's rank correlation rho

data:  spo2_VPO and hb
S = 2274841, p-value = 0.3201
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.06527711
```

SpO2 and hemoglobin were not correlated (rho= -0.065, p=0.32).

Sex and sleep apnea

Mean expected frequency:

```
  mean_expected_freq
1                 59
```

Since value is grater than 5.0, chi-squared without continuity correction is appropriate.

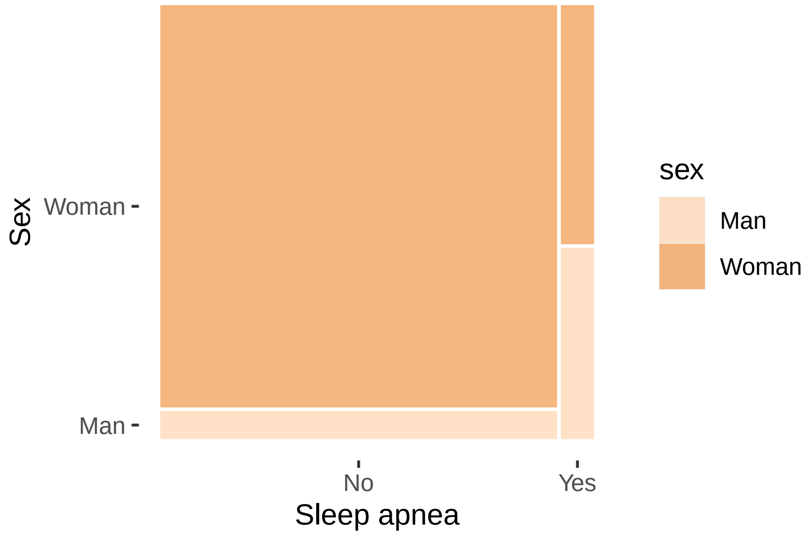
Frequencies:

```
       sleep_apnea
sex      No Yes
  Man    14   8
  Woman 204  10
```

Percentage:

```
       sleep_apnea
sex       No  Yes
  Man   63.6 36.4
  Woman 95.3  4.7
```

Mosaic Plot



```
    Pearson's Chi-squared test

data:  frequencies
X-squared = 28.437, df = 1, p-value = 9.68e-08
```

Sex was associated with OSA (p<0.001) as men had the diagnosis more frequently compared to women.

Package References

- Fox J, Weisberg S (2019). *An R Companion to Applied Regression*, Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.
- Fox J, Weisberg S, Price B (2022). *carData: Companion to Applied Regression Data Sets*. R package version 3.0-5, https://CRAN.R-project.org/package=carData.
- Grolemund G, Wickham H (2011). "Dates and Times Made Easy with lubridate." *Journal of Statistical Software*, *40*(3), 1-25. https://www.jstatsoft.org/v40/i03/.
- Jeppson H, Hofmann H, Cook D (2021). *ggmosaic: Mosaic Plots in the 'ggplot2' Framework*. R package version 0.3.3, https://CRAN.R-project.org/package=ggmosaic.
- Makowski D, Lüdecke D, Patil I, Thériault R, Ben-Shachar M, Wiernik B (2023). "Automated Results Reporting as a Practical Tool to Improve Reproducibility and Methodological Best Practices Adoption." *CRAN*. https://easystats.github.io/report/.
- Müller K, Wickham H (2023). *tibble: Simple Data Frames*. R package version 3.2.1, https://CRAN.R-project.org/package=tibble.
- Pinheiro J, Bates D, R Core Team (2023). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-164, https://CRAN.R-project.org/package=nlme. Pinheiro JC, Bates DM (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York. doi:10.1007/b98882 https://doi.org/10.1007/b98882.
- R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
- Rich B (2023). *table1: Tables of Descriptive Statistics in HTML*. R package version 1.4.3, https://CRAN.R-project.org/package=table1.
- Rinker TW, Kurkiewicz D (2018). *pacman: Package Management for R*. version 0.5.0, http://github.com/trinker/pacman.
- Textor J, van der Zander B, Gilthorpe MS, Liśkiewicz M, Ellison GT (2016). "Robust causal inference using directed acyclic graphs: the R package 'dagitty'." *International Journal of Epidemiology*, *45*(6), 1887-1894. doi:10.1093/ije/dyw341 https://doi.org/10.1093/ije/dyw341.
- Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.
- Wickham H (2023). *forcats: Tools for Working with Categorical Variables (Factors)*. R package version 1.0.0, https://CRAN.R-project.org/package=forcats.
- Wickham H (2023). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.5.1, https://CRAN.R-project.org/package=stringr.
- Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K,

Yutani H (2019). “Welcome to the tidyverse.” *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686.

- Wickham H, François R, Henry L, Müller K, Vaughan D (2023). *dplyr: A Grammar of Data Manipulation.* R package version 1.1.4, https://CRAN.R-project.org/package=dplyr.
- Wickham H, Henry L (2023). *purrr: Functional Programming Tools.* R package version 1.0.2, https://CRAN.R-project.org/package=purrr.
- Wickham H, Hester J, Bryan J (2023). *readr: Read Rectangular Text Data.* R package version 2.1.4, https://CRAN.R-project.org/package=readr.
- Wickham H, Vaughan D, Girlich M (2023). *tidyr: Tidy Messy Data.* R package version 1.3.0, https://CRAN.R-project.org/package=tidyr.
- Wood SN (2011). “Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models.” *Journal of the Royal Statistical Society (B)*, *73*(1), 3-36. Wood S, N., Pya, S”afken B (2016). “Smoothing parameter and model selection for general smooth models (with discussion).” *Journal of the American Statistical Association*, *111*, 1548-1575. Wood SN (2004). “Stable and efficient multiple smoothing parameter estimation for generalized additive models.” *Journal of the American Statistical Association*, *99*(467), 673-686. Wood S (2017). *Generalized Additive Models: An Introduction with R*, 2 edition. Chapman and Hall/CRC. Wood SN (2003). “Thin-plate regression splines.” *Journal of the Royal Statistical Society (B)*, *65*(1), 95-114.