Preoperative Atelectasis

Part 6: Statistical Modelling of SpO2

Javier Mancilla Galindo

2024-04-17

Table of contents

Setup

Packages used

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


pacman::p_load(
  tidyverse, # Used for basic data handling and visualization.
  RColorBrewer, #Color palettes for data visualization.
  table1, #Used to add lables to variables.
  gridExtra, #Used to arrange multiple ggplots in a grid.
  grid, #Used to arrange multiple ggplots in a grid.
  CBPS, #Used to calculate non-parametric propensity scores for IPW.
  WeightIt, #Used to calculate weights from propensity scores for IPW.
  mgcv, #Used to model non-linear relationships with a general additive model.
  gt, #Used to present a summary of the results of tables.
  gratia, #Used together with gglopt2 to create smooth partial effects plot
        # from gam models.
  metR, # Used to plot predictions of SpO2.
  report #Used to cite packages used in this session.
)
```

Session and package dependencies

```
R version 4.3.3 (2024-02-29 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 11 x64 (build 22631)

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] grid      stats     graphics  grDevices datasets  utils     methods
[8] base

other attached packages:
 [1] report_0.5.8        metR_0.15.0         gratia_0.8.2
 [4] gt_0.10.1           mgcv_1.9-1          nlme_3.1-164
 [7] WeightIt_1.0.0      CBPS_0.23           glmnet_4.1-8
[10] Matrix_1.6-5        numDeriv_2016.8-1.1 nnet_7.3-19
[13] MatchIt_4.5.5       MASS_7.3-60.0.1     gridExtra_2.3
[16] table1_1.4.3        RColorBrewer_1.1-3  lubridate_1.9.3
[19] forcats_1.0.0       stringr_1.5.1       dplyr_1.1.4
[22] purrr_1.0.2         readr_2.1.5         tidyr_1.3.1
[25] tibble_3.2.1        ggplot2_3.5.0       tidyverse_2.0.0
[28] pacman_0.5.1
```

Model SpO2

The SpO2 variable does not have a normal distribution. Furthermore, the distance between 1% increases in SpO2 cannot be considered equidistant increases since values are determined from the S-shaped curve of hemoglobin saturation. This is the reason why the distribution of SpO2 is negatively skewed, with upper values reaching the saturation point of the hemoglobin curve.

Therefore, modelling SpO2 as a linear term could be potentially misleading. Nonetheless, a model assuming a gaussian distribution for SpO2 may potentially be easier to understand and communicate.

Thus, I first created model SpO2 assuming a gaussian distribution and then applied a fractional regression model which is more appropriate for the distribution of this variable. The rationale for this was that if conclusions were not different with both models, presenting a model assuming a gaussian distribution would have been easier to understand and communicate. However, since conclusions were indeed different, I will present the results for the more appropriate fractional regression model.

As a last note, I first assessed the relationships between variables without removing any outliers. Examination of residuals showed that there were some influential outliers having an impact on the models. Thus, I decided to remove a total of 7 outliers only for the SpO2 models shown here (3 for the SpO2 ~ BMI relationship and 4 for SpO2 ~ atelectasis percent). This document presents the results of analyses after removing outliers. This code can be readapted to run all analyses without the removal of outliers.

Fractional regression model

Convert SpO2 to fractional values between 0 and 1 to model.

```r
data <- data %>% mutate(spo2_fraction = spo2_VPO/100)
```

Empty model

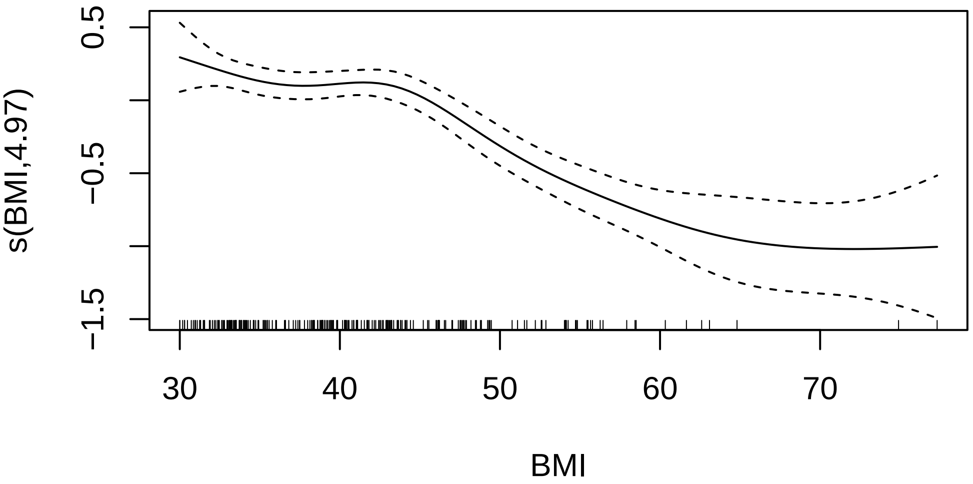
First, I will fit an empty model

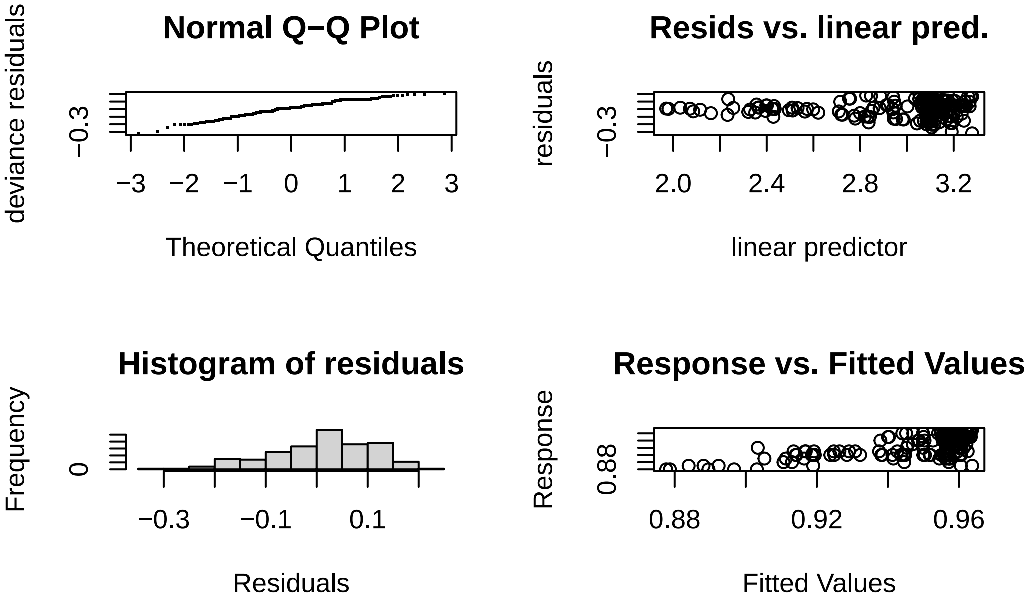
BMI smooth term and residuals

Model with a smooth BMI term as the only explanatory variable.

Since we are now using a different family function (quasibinomial with logit link) and we are no longer assuming a Gaussian distribution (which was done in figure 1 for an initial impression of the relationship between variables), it is important to determine the k value that offers the best representation of the change in the outcome variable with this function. I checked this by varying the value of k in the following code and ***k=8**** offered the best visual representation with the largest increase in deviance explained and optimal k-index. While varying k, it will be noted that the highest edf occurs at k=8. This can be replicated by varying the value of k in the following code:

```
       k'      edf  k-index p-value
s(BMI)  7 4.967787 1.102615  0.9425


Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ s(BMI, k = 8)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.98458    0.03212   92.92   <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(BMI) 4.968  5.829 21.79  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.386   Deviance explained = 34.1%
GCV = 0.010718  Scale est. = 0.010826  n = 236
```





```
Method: GCV   Optimizer: outer newton
full convergence after 4 iterations.
Gradient range [-4.075644e-09,-4.075644e-09]
```

```
(score 0.01071767 & scale 0.0108264).
Hessian positive definite, eigenvalue range [2.025108e-05,2.025108e-05].
Model rank =  8 / 8

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

         k'  edf k-index p-value
s(BMI) 7.00 4.97     1.1    0.94
```

There are influential residuals. Will assess which of these could be removed according to Cook's distance.

```r
data %>%
  mutate(
  cooksd = cooks.distance(model_BMI),
  outlier = ifelse(cooksd < 4/nrow(data), "keep","delete")
) %>%
  filter(outlier=="delete") %>%
  dplyr::select(ID,BMI,spo2_VPO,cooksd,outlier) %>%
  arrange(desc(cooksd)) %>%
  gt()
```

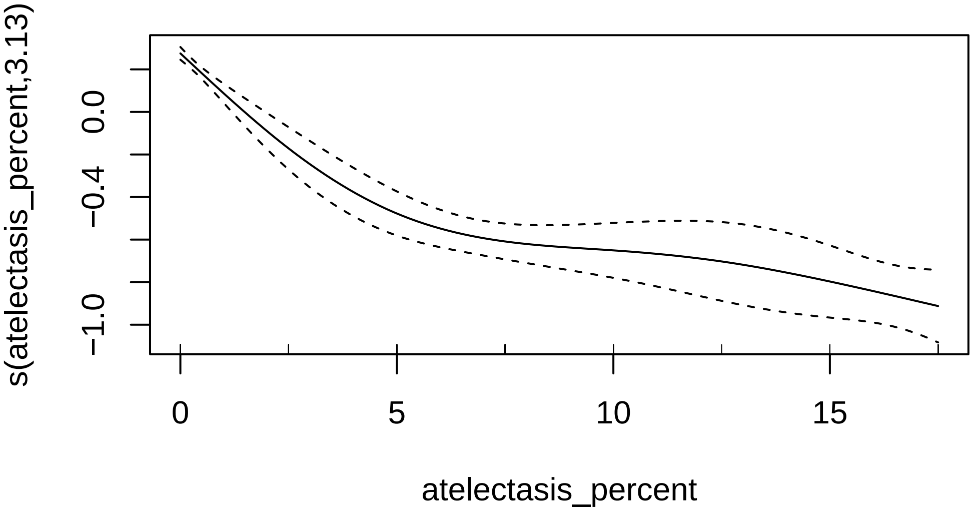
ID	Body mass index	Oxygen saturation (SpO2)	Oxygen saturation (SpO2)	outlier
122	30.00	89	0.13348733	delete
140	32.46	89	0.03196280	delete
102	30.00	99	0.01698986	delete

Now, I will examine atelectasis percentage:

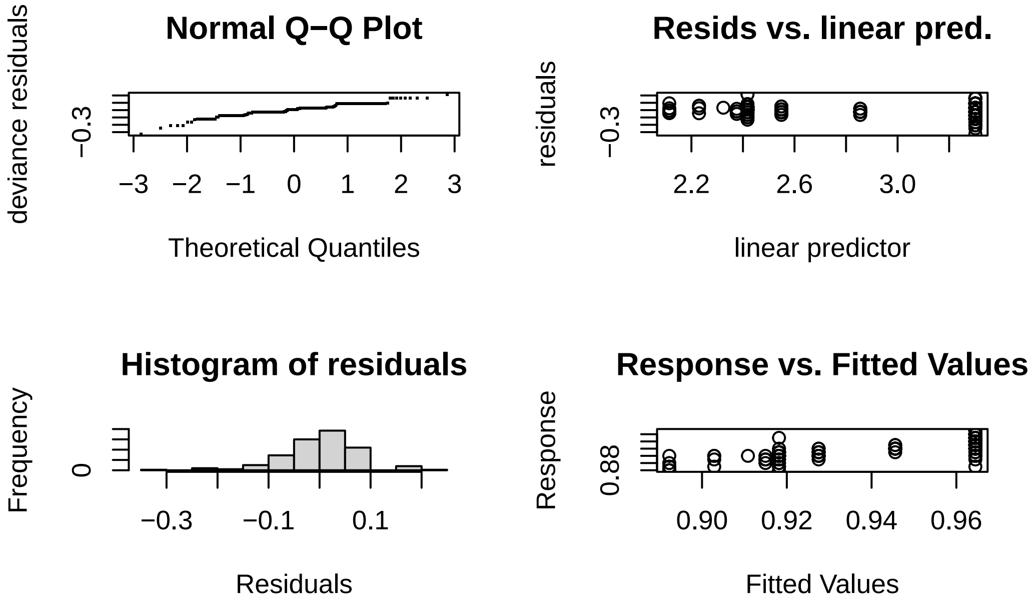
Atelectasis percent smooth term and residuals

Using a smooth term for atelectasis percentage is comparable to having it as categorical, which can be checked by substituting the smooth term for the categorical term in the models. However, the smooth term will allow to have a visual representation of the partial effect of atelectasis percent on SpO2 and compare it to BMI, which is why I decided to keep the smooth term to model the effect of atelectasis.

I determined the optimal k value for atelectasis at k=5. A term with a lower k (i.e., k=3) lead to a decrease in explained deviance compared to the categorical term and was not a good representation of the trend in values for the variable, especially at the higher values as it curved upwards compared to the downward trend seen. Therefore, I kept k=5.



```
Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ s(atelectasis_percent, k = 5)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.02654    0.02526   119.8   <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(atelectasis_percent) 3.131  3.592 105.3  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.676   Deviance explained = 62.3%
GCV = 0.0060413  Scale est. = 0.0062171  n = 236
```



```
Method: GCV   Optimizer: outer newton
full convergence after 3 iterations.
Gradient range [1.135994e-09,1.135994e-09]
(score 0.006041257 & scale 0.006217078).
Hessian positive definite, eigenvalue range [2.484316e-05,2.484316e-05].
Model rank =  5 / 5

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(atelectasis_percent) 4.00 3.13    1.06    0.82
```

Outliers:

```
data %>%
  mutate(
  cooksd = cooks.distance(model_atel_smooth),
  outlier = ifelse(cooksd < 4/nrow(data), "keep","delete")
) %>%
  filter(outlier=="delete") %>%
  dplyr::select(ID,BMI,spo2_VPO,atelectasis_percent,cooksd,outlier) %>%
arrange(desc(cooksd)) %>%
gt()
```

ID	Body mass index	Oxygen saturation (SpO2)	Percentage of atelectasis	Oxygen saturation (SpO2)
205	55.44	92	17.5	0.04401798
122	30.00	89	0.0	0.03928085
168	48.78	97	7.5	0.03216437
114	41.10	91	0.0	0.02101897
103	60.33	88	7.5	0.01743817

I will remove influential outliers with cooks distance >0.05 for BMI:

Likewise, remove influential outliers with cooks distance >0.05 for atelectasis percent:

Inverse probability weighting

In order to account for exposure-outcome confounding and also mediator-outcome confounding, inverse probability weighting is a good modelling option since the mediator-outcome confounding affected by the exposure induces collider-stratification bias. Thus, a propensity score will be calculated for both the exposure (BMI) and mediator (atelectasis percent). Inverse probability weights will be obtained from propensity scores and combined into a single weight that will be used to model SpO2.

Propensity scores and weights

Non-Parametric Covariate Balancing Propensity Score (npcbps) will be obtained to avoid problems in the fitting of the models due to skewed distributions, thereby adopting a distribution-free method for obtaining the weights. Npcbps are directly interpretable as inverse probability weights (see CBPS documentation), reason why the additional step of obtaining inverse weights is not needed with this method.

Weights for exposure (BMI):

```
data$weight1 <- weightit(BMI ~ age + sex + altitude_cat,
                         data,
                         method = "npcbps",
                         over = FALSE)$weights
```

Weights for mediator (atelectasis percent):

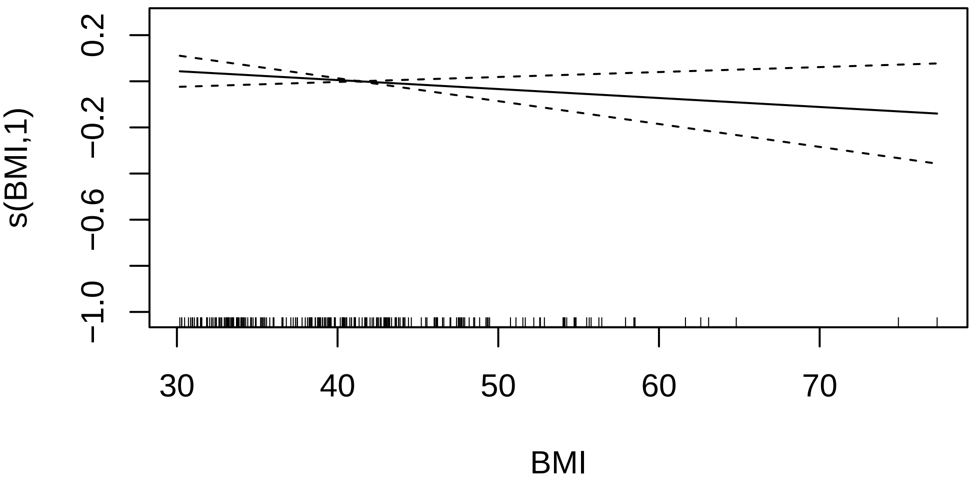
```
data$weight2 <- weightit(
  factor(atelectasis_percent, ordered = TRUE) ~
    BMI + age + sex + altitude_cat + asthma + sleep_apnea + COPD,
  data,
  method = "npcbps")$weights
```

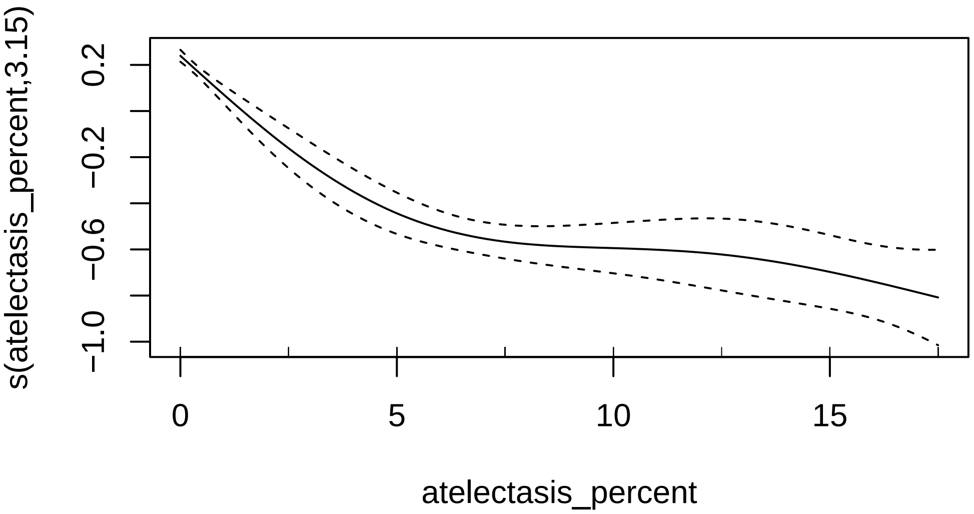
Overall weight:

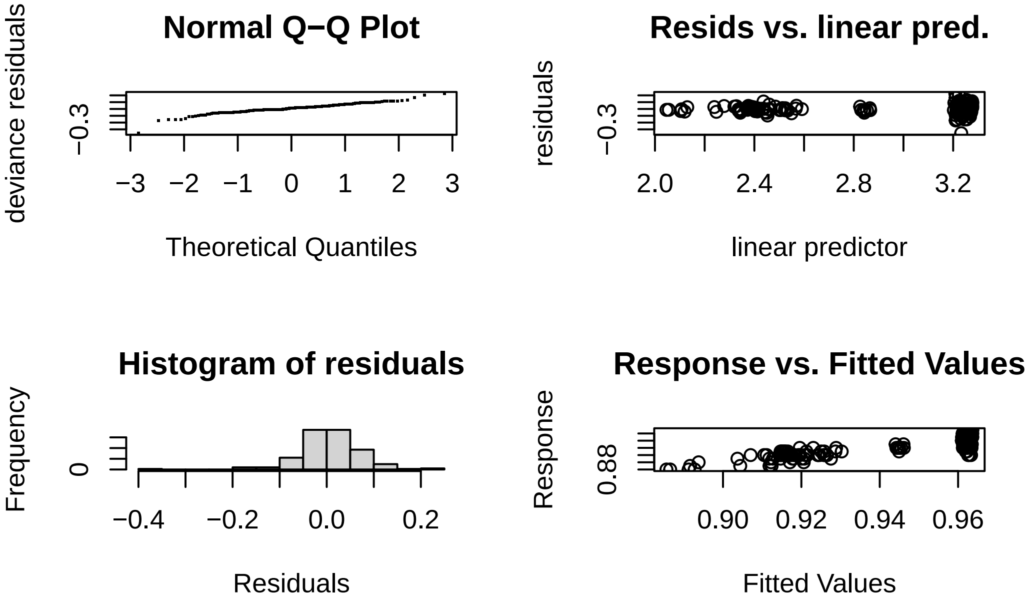
```
data <- data %>%
  mutate(
  weight = weight1*weight2
  )
```

Fit IPW model including both BMI and atelectasis percentage:

```
Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ s(BMI, k = 8) + s(atelectasis_percent, k = 5)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.99448    0.02166   138.2   <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(BMI)                 1.000  1.000  1.659   0.199
s(atelectasis_percent) 3.152  3.614 97.889  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.716   Deviance explained = 64.9%
GCV = 0.0043753  Scale est. = 0.0043973  n = 229
```







```
Method: GCV   Optimizer: outer newton
full convergence after 6 iterations.
Gradient range [-2.938139e-09,-2.517301e-10]
```

```
(score 0.004375297 & scale 0.004397271).
Hessian positive definite, eigenvalue range [2.936941e-09,1.920412e-05].
Model rank =  12 / 12

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                         k'  edf k-index p-value
s(BMI)                 7.00 1.00    1.09    0.95
s(atelectasis_percent) 4.00 3.15    1.03    0.76
```

Very influential outlier.

```
data %>%
  mutate(
  cooksd = cooks.distance(model_plus),
  outlier = ifelse(cooksd < 4/nrow(data), "keep","delete")
) %>%
  filter(outlier=="delete") %>%
  dplyr::select(ID,BMI,spo2_VPO,atelectasis_percent,cooksd,outlier) %>%
  arrange(desc(cooksd)) %>%
  gt()
```

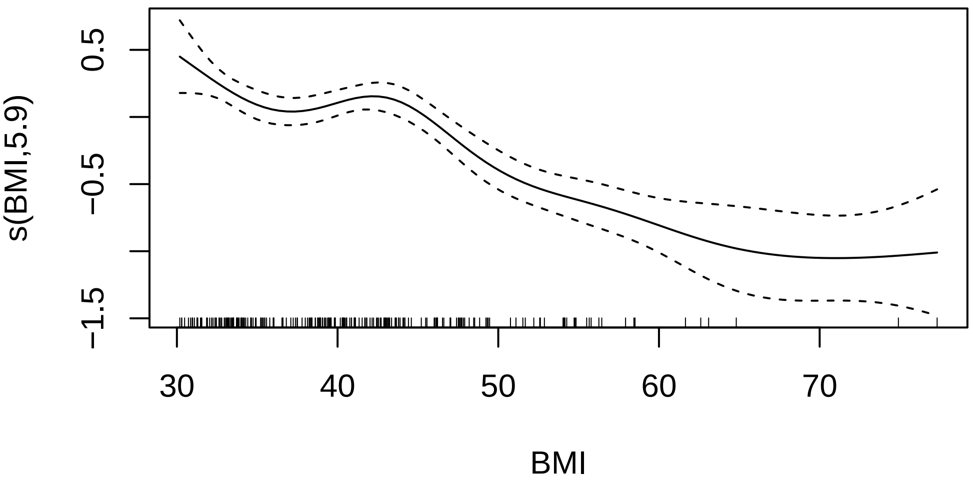
ID	Body mass index	Oxygen saturation (SpO2)	Percentage of atelectasis	Oxygen saturation (SpO2)
107	41.74	94	0.0	0.55248730
165	34.93	91	7.5	0.13631966
20	47.83	98	0.0	0.08991243
39	47.51	94	0.0	0.03994338
40	33.11	93	7.5	0.03639909
200	39.05	94	7.5	0.02788647
12	46.19	94	0.0	0.02575926
18	47.55	98	0.0	0.02431717
54	34.60	96	0.0	0.02208772
58	52.87	92	12.5	0.01847974

I will remove this extreme outlier:

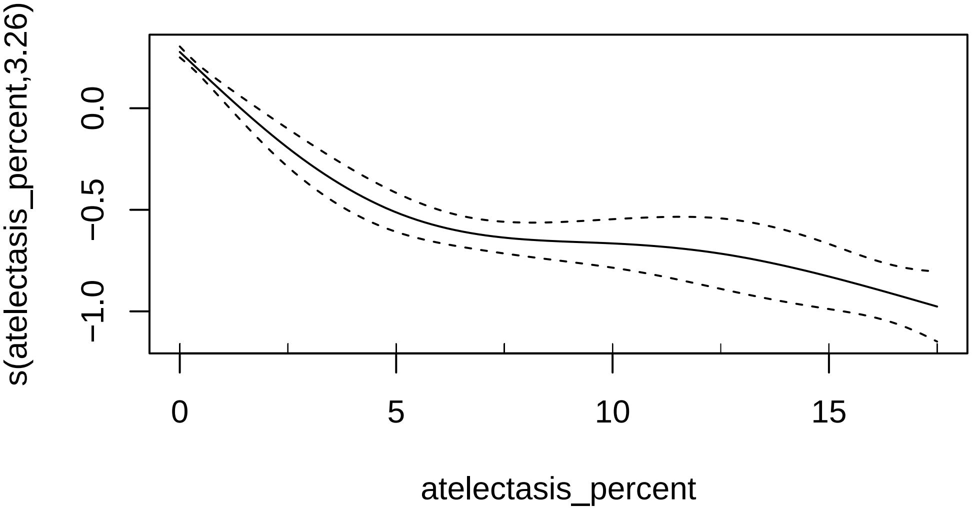
BMI only model (unadjusted, unweighted)

```
Family: quasibinomial
```

```
Link function: logit

Formula:
spo2_fraction ~ s(BMI, k = 8)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.00860    0.03096   97.17   <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(BMI) 5.901  6.587 22.5  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.425   Deviance explained = 38.3%
GCV = 0.0096285  Scale est. = 0.0094719  n = 228
```



Atelectasis percent model (unadjusted, unweighted)

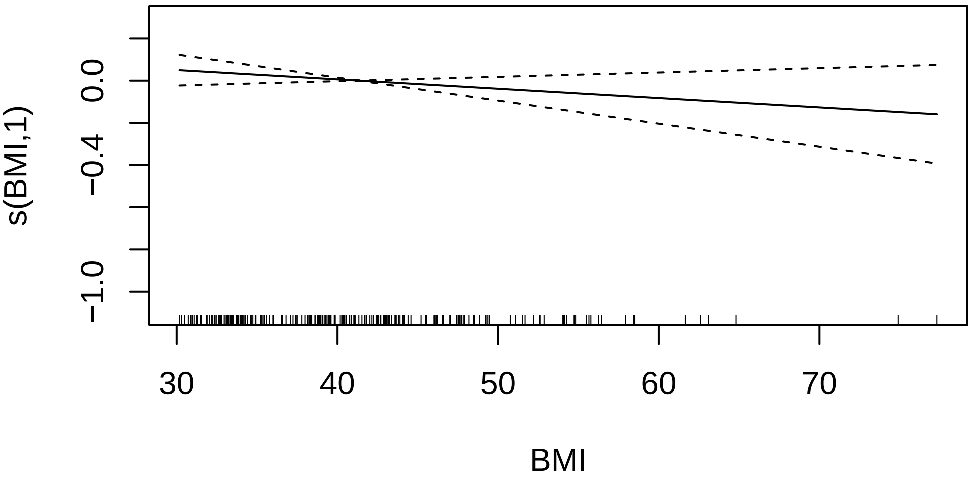


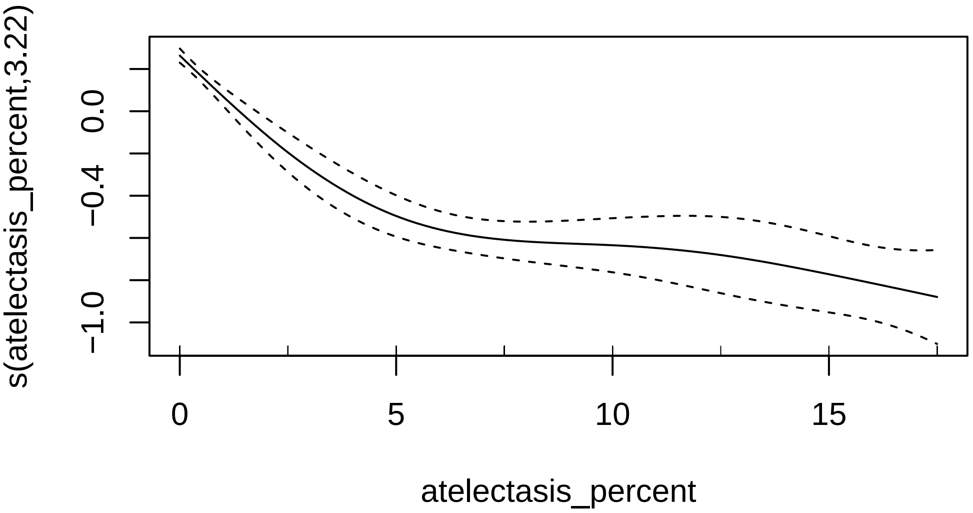
```
Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ s(atelectasis_percent, k = 5)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.04997    0.02321   131.4   <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(atelectasis_percent) 3.262  3.698 128.3  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.737   Deviance explained = 67.5%
GCV = 0.0049571  Scale est. = 0.004952  n = 228
```

s(BMI) + s(atelectasis percentage), unadjusted, unweighted

Fit model sBMI plus atelectasis percentage:

```

Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ s(BMI, k = 8) + s(atelectasis_percent, k = 5)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.05022    0.02318   131.6   <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(BMI)                 1.000  1.000  1.864   0.174
s(atelectasis_percent) 3.223  3.664 67.299  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =   0.74   Deviance explained = 67.7%
GCV = 0.004962  Scale est. = 0.0049392  n = 228
```

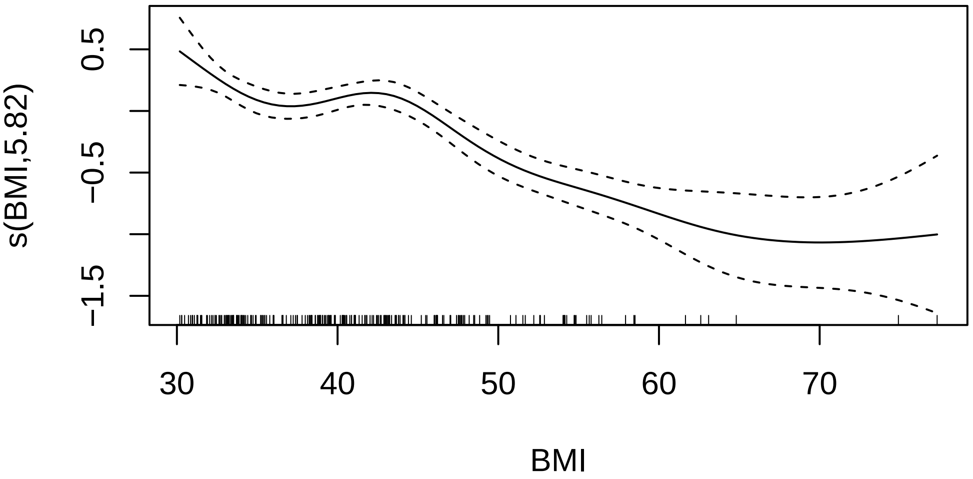

s(BMI,1)
0.0
−0.4
−1.0
30
40
50
60
70
BMI


s(atelectasis_percent,3.22)
0.0
−0.4
−1.0
0
5
10
15
atelectasis_percent

Adjusted model BMI

Fit IPW model for BMI. This model accounts for confounders relevant to BMI, after having obtained a balanced pseudopopulation through propensity weighting as detailed earlier.

```
Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ s(BMI, k = 8)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.00884    0.03099    97.1   <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(BMI) 5.816  6.538 22.32  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.421   Deviance explained = 38.2%
GCV = 0.0095751  Scale est. = 0.009433  n = 228
```

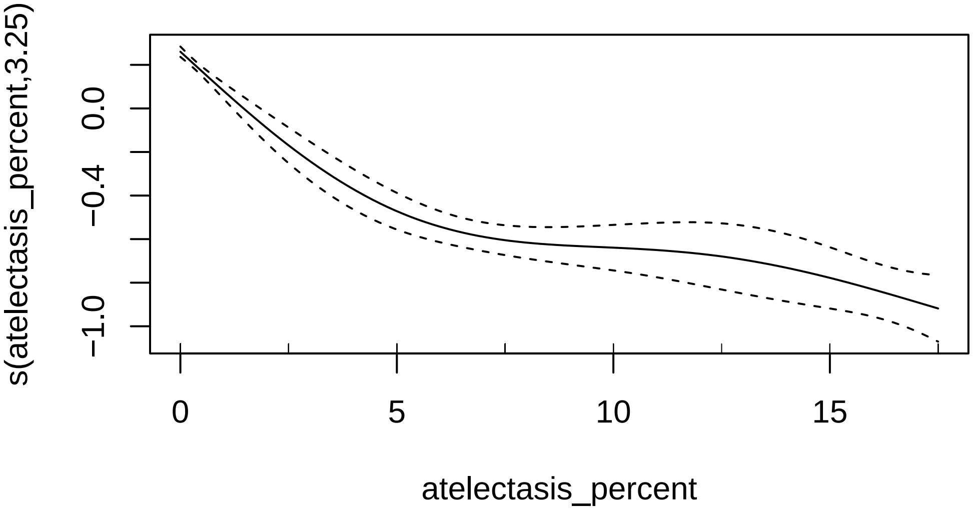


Adjusted model Atelectasis Percent

Fit IPS model for atelectasis percent:

```
Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ s(atelectasis_percent, k = 5)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.02035    0.02054     147   <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(atelectasis_percent) 3.25  3.688 146.7  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

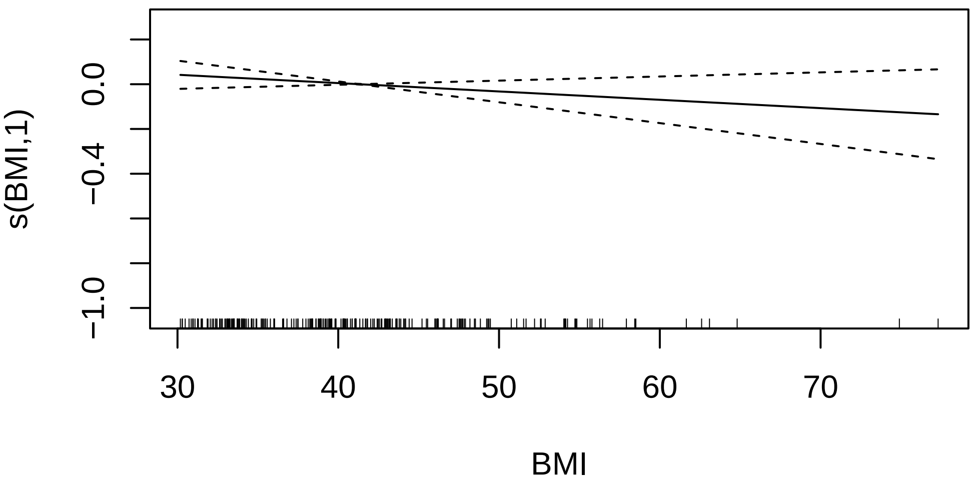
```
R-sq.(adj) =  0.762   Deviance explained = 70.2%
GCV = 0.0038206  Scale est. = 0.0037876  n = 228
```

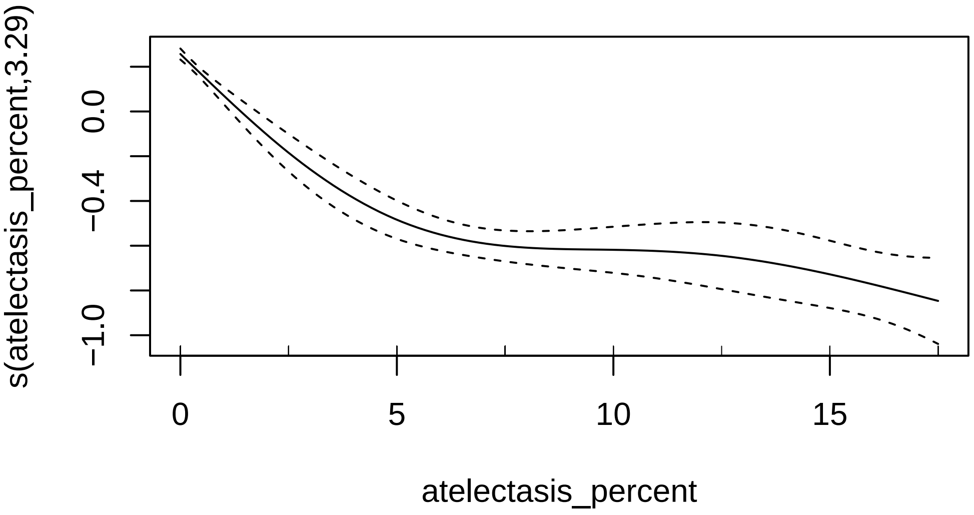


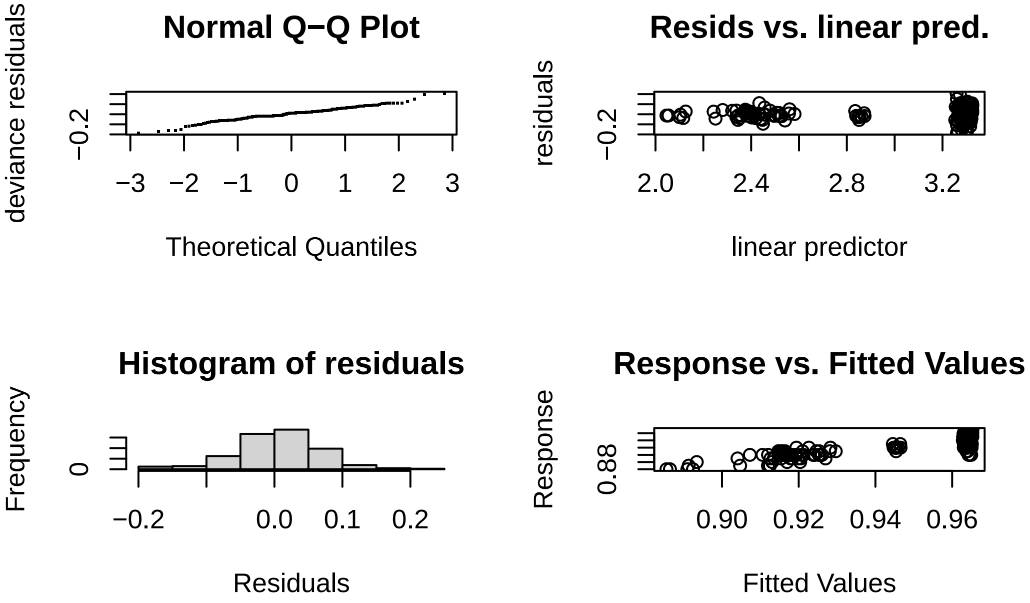
Fully adjusted model (IPW model)

```
Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ s(BMI, k = 8) + s(atelectasis_percent, k = 5)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.02676    0.02083   145.3   <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(BMI)                1.000  1.001   1.789   0.182
s(atelectasis_percent) 3.289  3.721 120.779  <2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.759   Deviance explained = 69.9%
GCV = 0.0037468  Scale est. = 0.0036974  n = 228
```







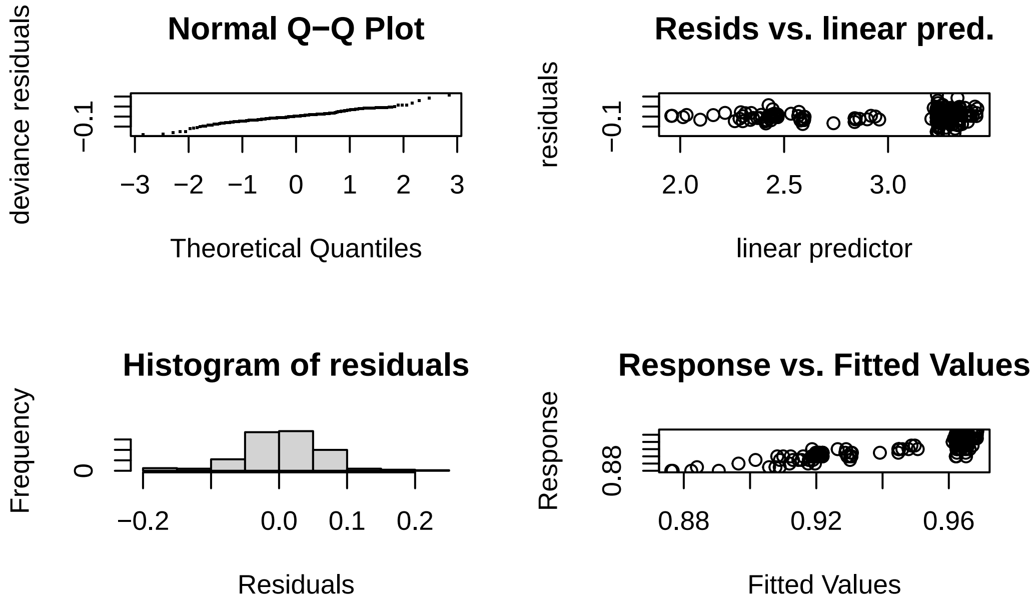
```
Method: GCV   Optimizer: outer newton
full convergence after 6 iterations.
Gradient range [-3.75327e-09,-3.839252e-10]
(score 0.003746773 & scale 0.003697431).
Hessian positive definite, eigenvalue range [3.752868e-09,1.62418e-05].
Model rank =  12 / 12

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                        k'  edf k-index p-value
s(BMI)                7.00 1.00    1.14    0.99
s(atelectasis_percent) 4.00 3.29    1.04    0.78
```

Note that there is complete separation of residuals, resembling oxygen categories. This is highly suggestive of this model not being good at explaining variation at higher oxygen levels (i.e. greater than 95%), whereas residuals are nearly randomly distributed at lower oxygen values. I will re-assess this hypothesis by the end of the document and in **Part 8**.

Test for interaction

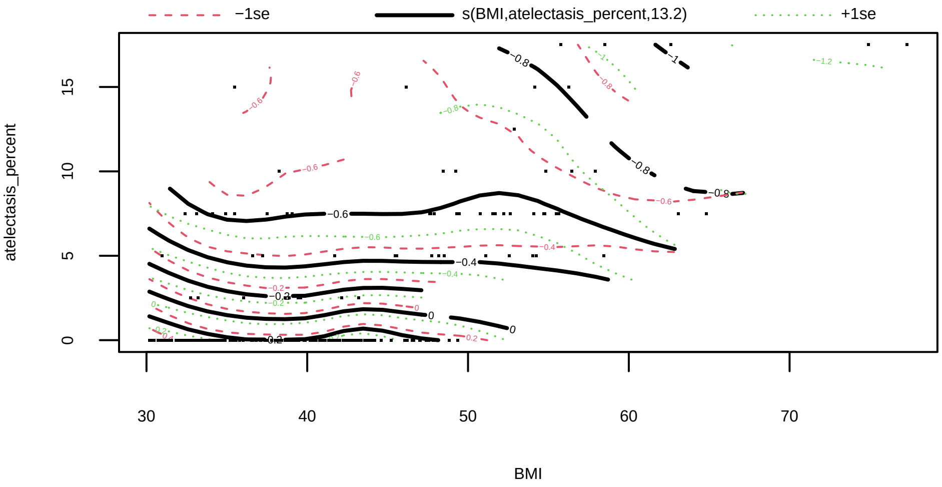
```
Family: quasibinomial
```

```
Link function: logit

Formula:
spo2_fraction ~ s(BMI, atelectasis_percent)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.03060    0.02315   130.9   <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(BMI,atelectasis_percent) 13.2  17.39 30.43  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.758   Deviance explained = 70.9%
GCV = 0.0039258  Scale est. = 0.0036899  n = 228
```

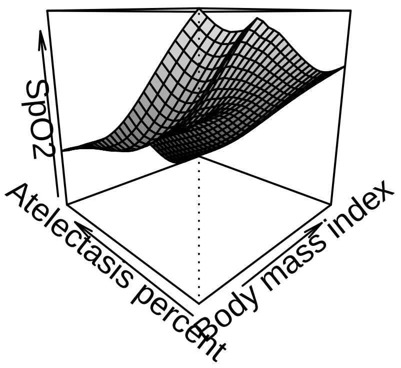


```
Method: GCV   Optimizer: outer newton
full convergence after 3 iterations.
```

```
Gradient range [1.048224e-10,1.048224e-10]
(score 0.003925793 & scale 0.003689942).
Hessian positive definite, eigenvalue range [9.732058e-05,9.732058e-05].
Model rank =  30 / 30

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                              k'  edf k-index p-value
s(BMI,atelectasis_percent) 29.0 13.2    1.07    0.94
```



Predicted SpO2



Not much improvement from a model with an interaction term. This model with an interaction term is very likely fitting noise.

Build a dataframe to compare models:

Models sorted by explained deviance (from higher to lower):

Model	aR2	dev
Interaction_adjusted	75.8	70.9
adjusted_atelectasis	76.2	70.2
Fully_adjusted	75.9	69.9
sBMI_atel	74.0	67.7
atel_smooth	73.7	67.5
sBMI	42.5	38.3
adjusted_BMI	42.1	38.2
empty	0.0	0.0

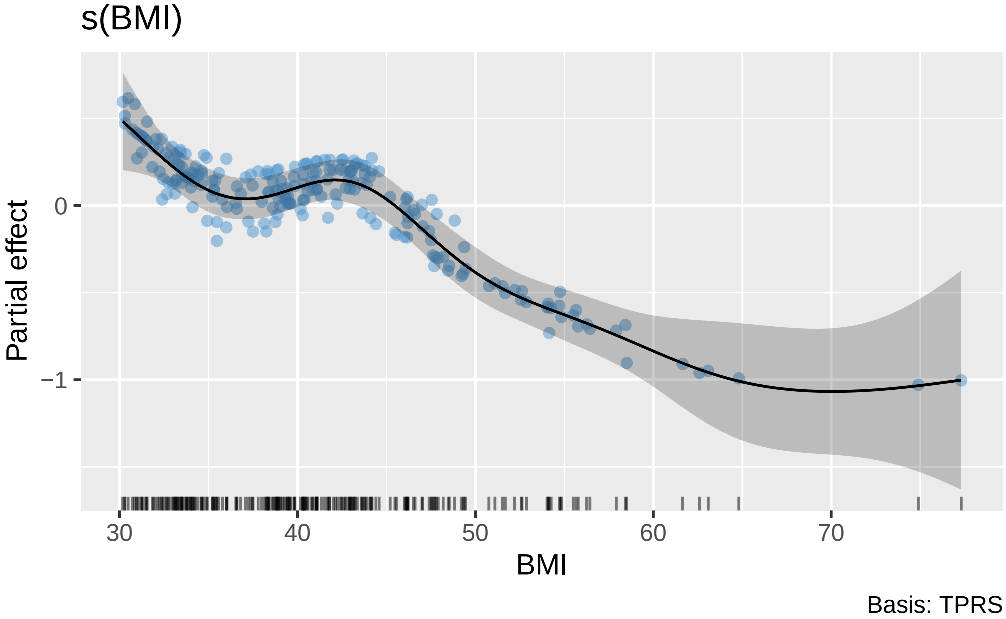
Model	aR2	dev
adjusted_atelectasis	76.2	70.2
Fully_adjusted	75.9	69.9
Interaction_adjusted	75.8	70.9
sBMI_atel	74.0	67.7

atel_smooth	73.7	67.5
sBMI	42.5	38.3
adjusted_BMI	42.1	38.2
empty	0.0	0.0

Figure SpO2 models

Figure 2a: Total effect of BMI (adjusted)

Assessment of residuals. This was done for all models.



Now, take the inverse logit function to assess partial effect on mean SpO2.

Partial effect on mean SpO2:

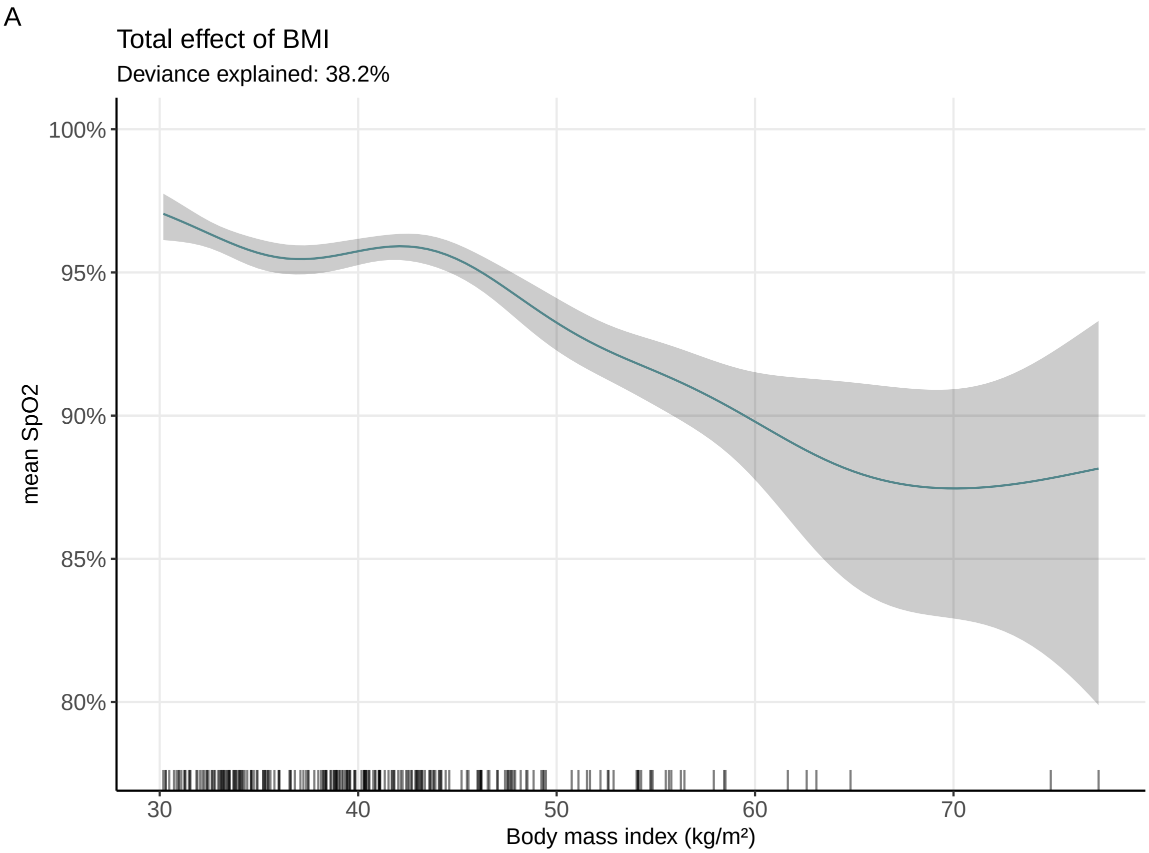


Figure 2b: Fully adjusted BMI

Partial effect on mean SpO2:

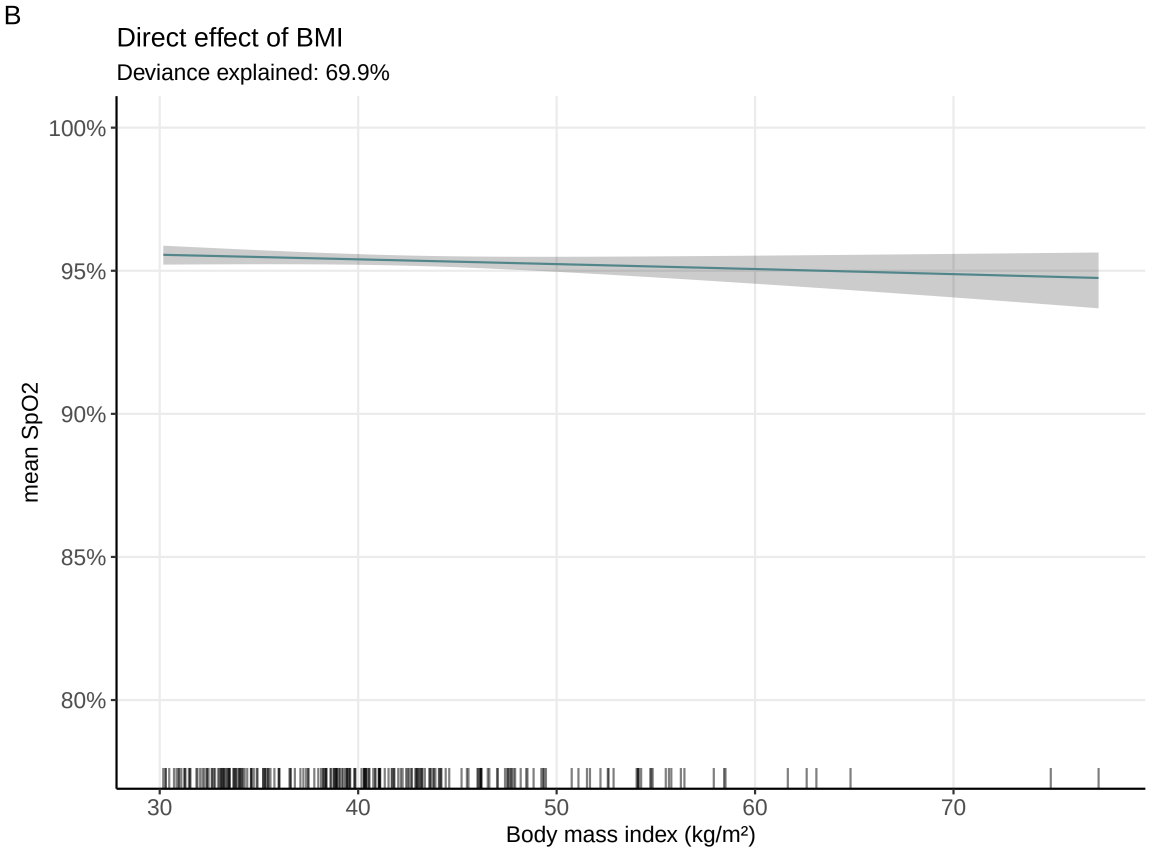


Figure 2c: sAtelectasis percent

Check residuals:

Draw a personalized plot:

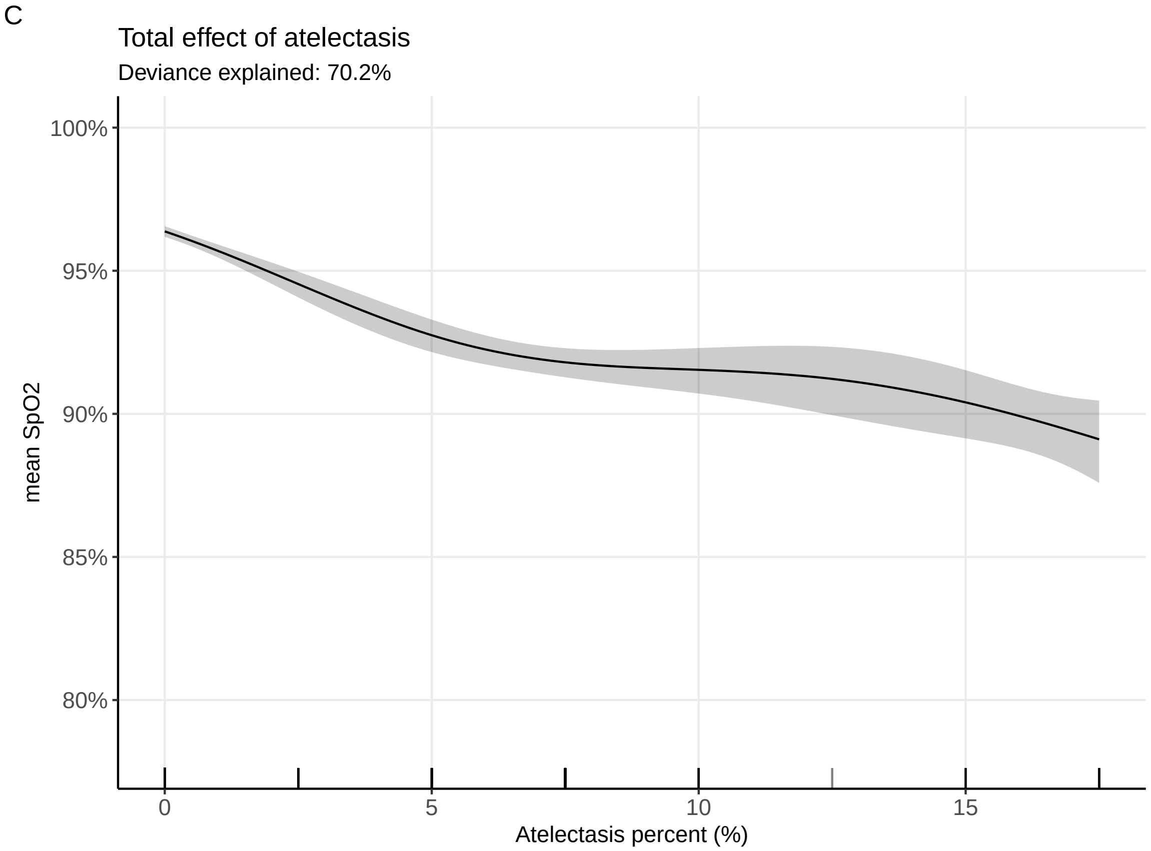


Figure 2d: Fully adjusted Atelectasis Percent

Partial effect on mean SpO2:

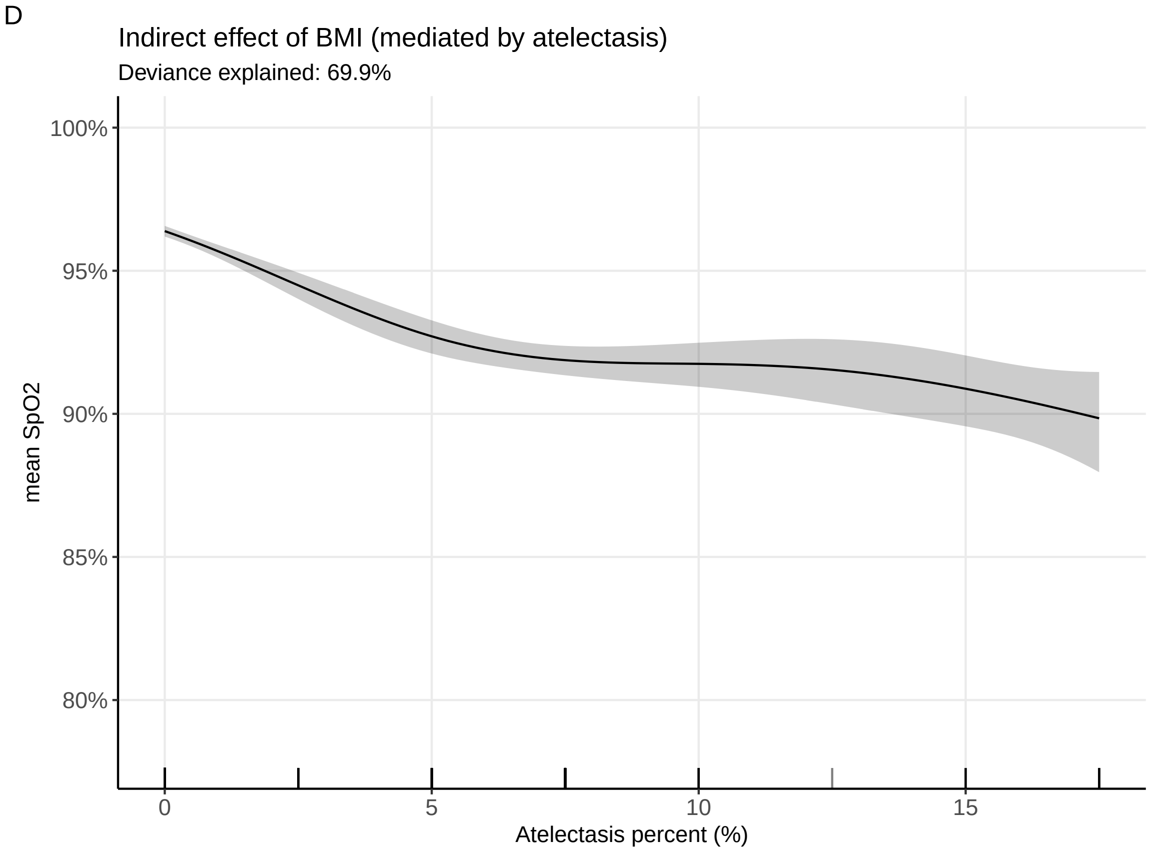
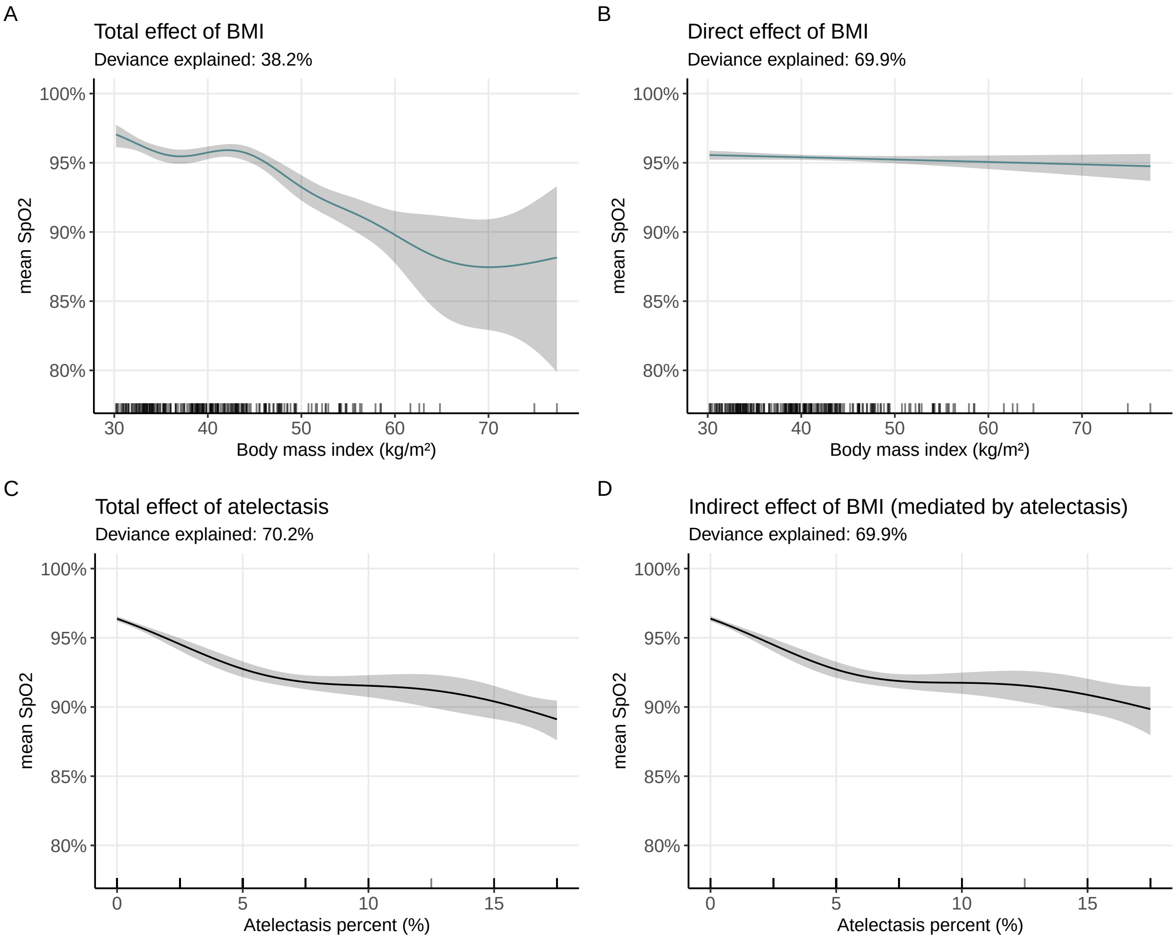

D
Indirect effect of BMI (mediated by atelectasis)
Deviance explained: 69.9%
mean SpO2
100%
95%
90%
85%
80%
0
5
10
15
Atelectasis percent (%)

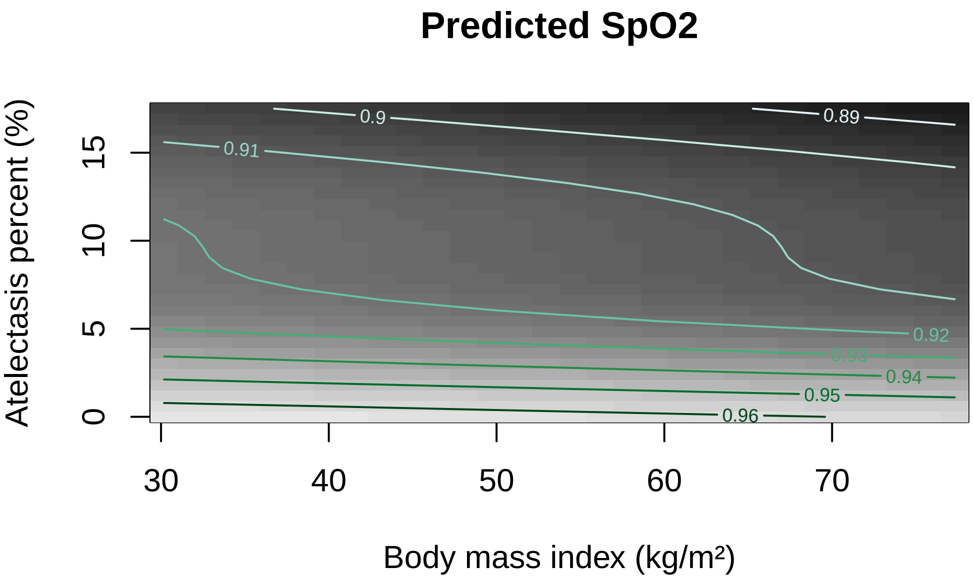
Figure 2


A
Total effect of BMI
Deviance explained: 38.2%
mean SpO2
100%
95%
90%
85%
80%
30
40
50
60
70
Body mass index (kg/m²)
B
Direct effect of BMI
Deviance explained: 69.9%
mean SpO2
100%
95%
90%
85%
80%
30
40
50
60
70
Body mass index (kg/m²)
C
Total effect of atelectasis
Deviance explained: 70.2%
mean SpO2
100%
95%
90%
85%
80%
0
5
10
15
Atelectasis percent (%)
D
Indirect effect of BMI (mediated by atelectasis)
Deviance explained: 69.9%
mean SpO2
100%
95%
90%
85%
80%
0
5
10
15
Atelectasis percent (%)

Predictions SpO2

These are the predicted SpO2 values in the fully adjusted model (adjusted_plus):

```r
vis.gam(model_plus,
        view=c("BMI","atelectasis_percent"),
        color = "gray",
        type = "response",
        plot.type = "contour",
        contour.col = brewer.pal(9, "BuGn"),
        nlevels=10,
        ylab = "Atelectasis percent (%)",
        xlab = "Body mass index (kg/m²)",
        main = "Predicted SpO2"
        )
```



This figure shows that this model is not able to predict SpO2 values above 96%. The range of predicted values of SpO2 that can be predicted with the data and model created in this study are within 88-96%. Lines correspond to a level of SpO2, so it can be seen that most of these are almost perpendicular to the y axis, meaning that most of the decrease in SpO2 is driven by increasing atelectasis percentage. Nonetheless, lines are not perfectly horizontal, which reflects that there is some residual effect of BMI on SpO2. Furthermore, this model shows that drops in SpO2 are more accentuated at the lower part of atelectasis percentage extension (93% to 96% mostly occur at atelectasis percentage lower than 5%). At SpO2 92% and lower, jumps are not as accentuated and there is a greater effect of increasing BMI as the lines tend to be more inclined. A 3D plot could perhaps allow to visualize these patterns if this is not clear enough from the 2D plot:

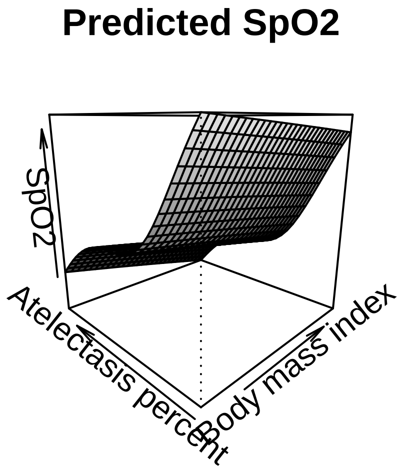
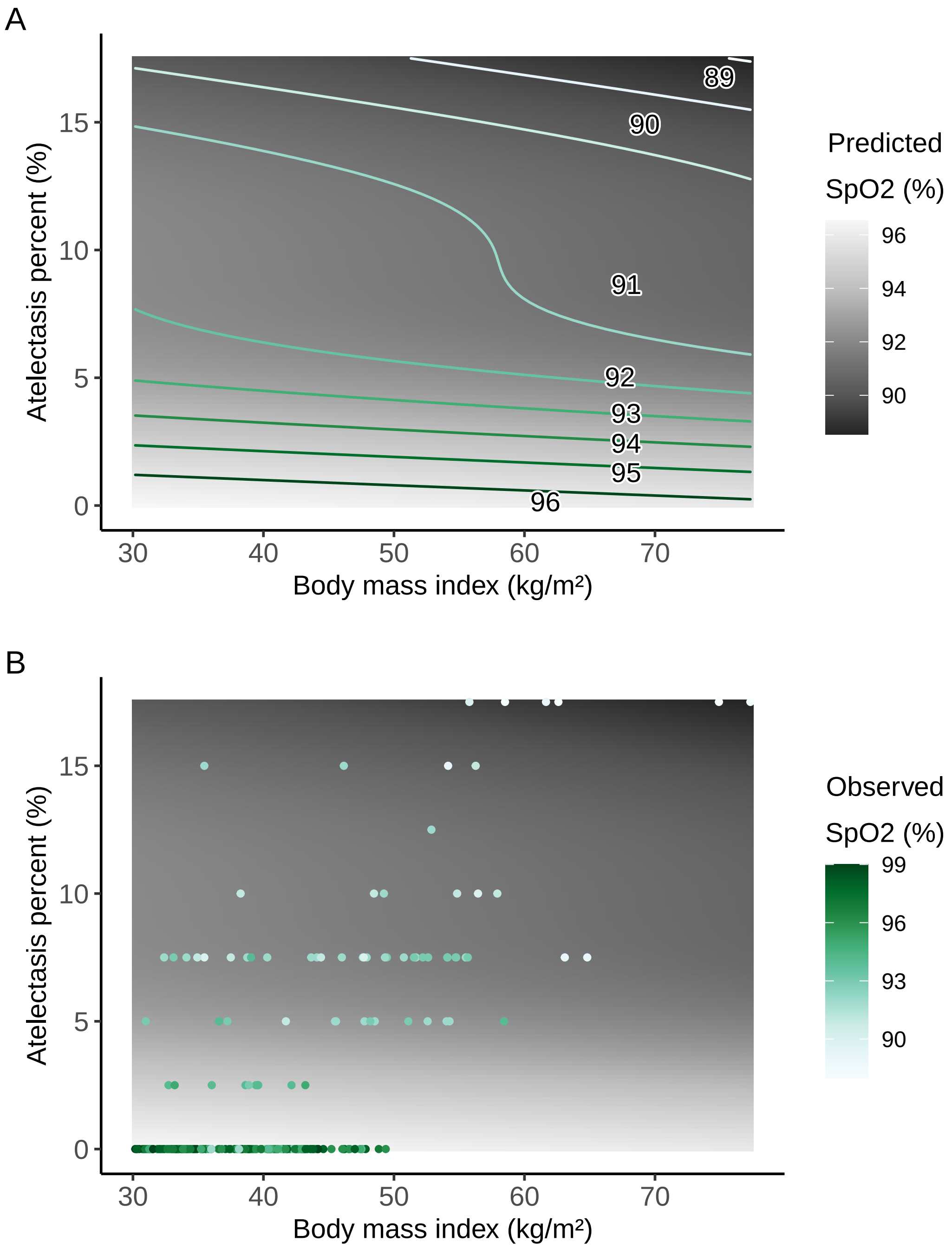


Figure 3

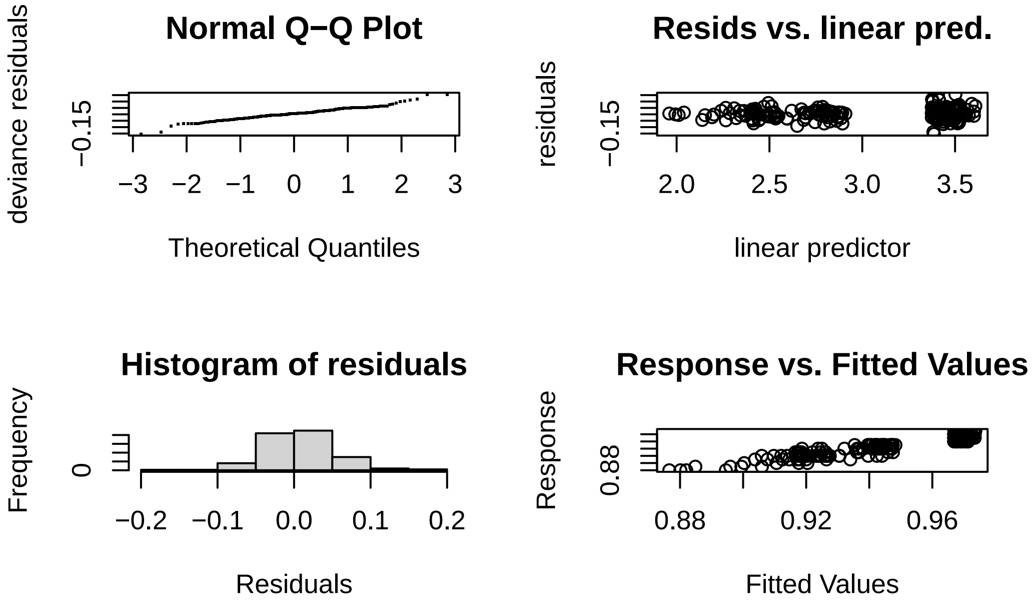
The 2D plot was recreated with the accompanying sourced script ***Figure3.R*** which also saves the 3D plot as FigureS5.


A
Atelectasis percent (%)
15
10
5
0
30
40
50
60
70
Body mass index (kg/m²)
89
90
91
92
93
94
95
96
Predicted
SpO2 (%)
96
94
92
90
B
Atelectasis percent (%)
15
10
5
0
30
40
50
60
70
Body mass index (kg/m²)
Observed
SpO2 (%)
99
96
93
90

Modelling subsets of low vs high SpO2

As mentioned earlier, there is complete separation of residuals, resembling oxygen categories. Perhaps including these in the model could allow to know the effect of atelectasis and BMI on SpO2 according to different SpO2 categories:

```
Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ s(BMI, k = 8) + s(atelectasis_percent, k = 5) +
    spo2_cat

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.34042    0.02709  123.32   <2e-16 ***
spo2_cat 95 -0.63095    0.04104  -15.37   <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(BMI)                6.192  6.774  3.163 0.00381 **
s(atelectasis_percent) 2.152  2.571 24.851 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.875   Deviance explained = 84.5%
GCV = 0.0020258  Scale est. = 0.0018961  n = 228
```



```
Method: GCV   Optimizer: outer newton
full convergence after 9 iterations.
Gradient range [2.298414e-11,5.759827e-11]
(score 0.002025796 & scale 0.001896135).
Hessian positive definite, eigenvalue range [5.390903e-06,9.319027e-06].
Model rank =  13 / 13

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                        k'  edf k-index p-value
s(BMI)                7.00 6.19    1.08    0.95
s(atelectasis_percent) 4.00 2.15    1.06    0.93
```

I will therefore model separately by splitting the dataset into participants with SpO2 lower than or equal to 95 vs those with SpO2 higher than 95. These analyses will be presented in **Part 8**.

Package References

- Auguie B (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3, https://CRAN.R-project.org/package=gridExtra.
- Bates D, Maechler M, Jagan M (2024). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.6-5, https://CRAN.R-project.org/package=Matrix.
- Campitelli E (2021). *metR: Tools for Easier Analysis of Meteorological Fields*. doi:10.5281/zenodo.2593516 https://doi.org/10.5281/zenodo.2593516, R package version 0.15.0, https://eliocamp.github.io/metR/.
- Fong C, Ratkovic M, Imai K (2022). *CBPS: Covariate Balancing Propensity Score*. R package version 0.23, https://CRAN.R-project.org/package=CBPS.
- Friedman J, Tibshirani R, Hastie T (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, *33*(1), 1-22. doi:10.18637/jss.v033.i01 https://doi.org/10.18637/jss.v033.i01. Simon N, Friedman J, Tibshirani R, Hastie T (2011). "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." *Journal of Statistical Software*, *39*(5), 1-13. doi:10.18637/jss.v039.i05 https://doi.org/10.18637/jss.v039.i05. Tay JK, Narasimhan B, Hastie T (2023). "Elastic Net Regularization Paths for All Generalized Linear Models." *Journal of Statistical Software*, *106*(1), 1-31. doi:10.18637/jss.v106.i01 https://doi.org/10.18637/jss.v106.i01.
- Gilbert P, Varadhan R (2019). *numDeriv: Accurate Numerical Derivatives*. R package version 2016.8-1.1, https://CRAN.R-project.org/package=numDeriv.
- Greifer N (2024). *WeightIt: Weighting for Covariate Balance in Observational Studies*. R package version 1.0.0, https://CRAN.R-project.org/package=WeightIt.
- 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/.
- Ho D, Imai K, King G, Stuart E (2011). "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." *Journal of Statistical Software*, *42*(8), 1-28. doi:10.18637/jss.v042.i08 https://doi.org/10.18637/jss.v042.i08.
- Iannone R, Cheng J, Schloerke B, Hughes E, Lauer A, Seo J (2024). *gt: Easily Create Presentation-Ready Display Tables*. R package version 0.10.1, https://CRAN.R-project.org/package=gt.
- 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.
- Neuwirth E (2022). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-3, https://CRAN.R-project.org/package=RColorBrewer.
- 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 (2024). *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.
- Simpson G (2024). *gratia: Graceful ggplot-Based Graphics and Other Functions for GAMs Fitted using mgcv*. R package version 0.8.2, https://gavinsimpson.github.io/gratia/.
- Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.
- Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.
- 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 (2024). *readr: Read Rectangular Text Data*. R package version 2.1.5, https://CRAN.R-project.org/package=readr.
- Wickham H, Vaughan D, Girlich M (2024). *tidyr: Tidy Messy Data*. R package version 1.3.1, 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.