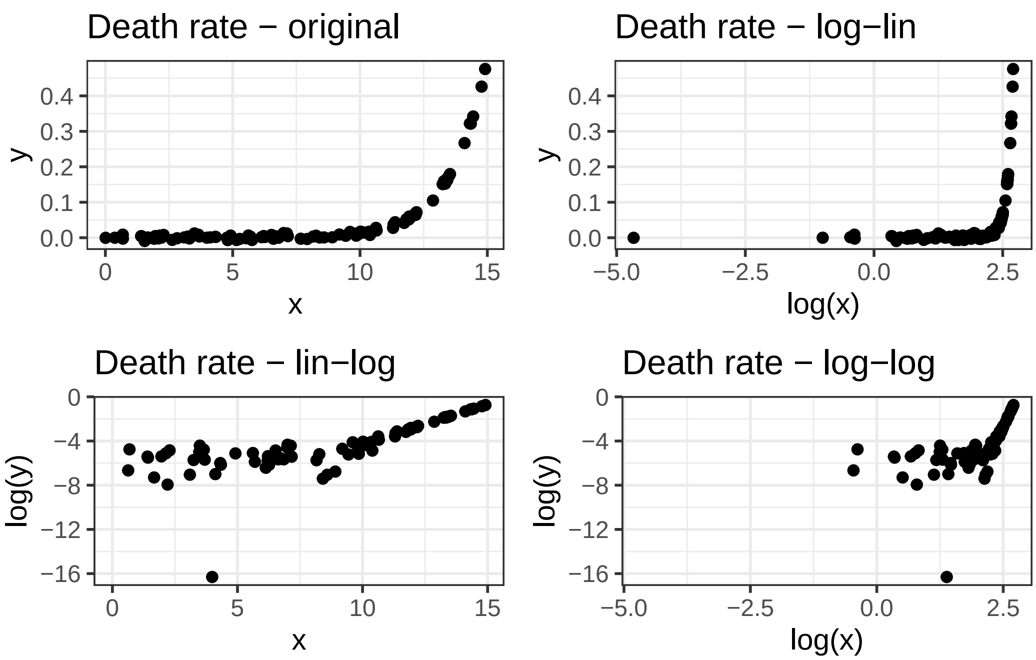
Recap exercises

Excurse: An inherently nonlinear model

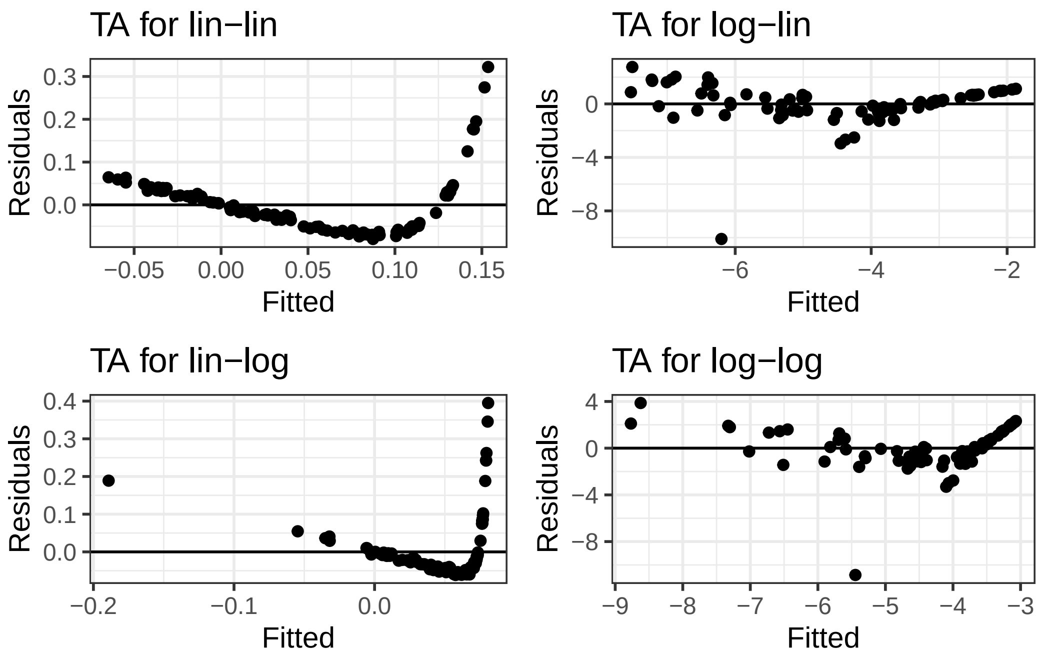
This is a model for death rates in certain situations:

$$y = \beta_0 + \lambda \exp\left(\beta_1^2 x\right) + \epsilon$$

It cannot be made linear in terms of parameters:



This can be verified using the TA plots:



Data wrangling

Read in the data set `wrangel_1.csv`. Transform the data set such that it can be considered tidy.

Then create a new data set that contains the means for all variables for each country. Missing values should be ignored when computing the means.

```
T4_df <- fread(here("data/wrangel_1.csv"), header = TRUE) %>%
  pivot_longer(
    cols = -all_of(c("country", "name")),
    names_to = "year",
    values_to = "value") %>%
  pivot_wider(names_from = "name", values_from = "value")
head(T4_df)
```

```
# A tibble: 6 x 5
  country year  Growth EducationSpending HealthSpending
  <chr>   <chr>  <dbl>             <dbl>          <dbl>
1 Germany 2005   0.732             NA              10.3
2 Germany 2006   3.82               4.29           10.2
3 Germany 2007   2.98               4.37           10.1
```

```
4 Germany 2008   0.960                  4.44           10.3
5 Germany 2009  -5.69                   4.91           11.2
6 Germany 2010   4.18                   4.94           11.1
```

```
T4_summary <- T4_df %>%
  group_by(country) %>%
  summarise(across(where(is.double), ~ mean(.x, na.rm=TRUE)))
T4_summary
```

```
# A tibble: 4 x 4
  country     Growth EducationSpending HealthSpending
  <chr>        <dbl>             <dbl>          <dbl>
1 Germany      1.17               4.79          11.0
2 Italy       -0.528              4.19           8.73
3 Netherlands  1.15               5.29          10.0
4 Spain        0.479              4.41           8.90
```

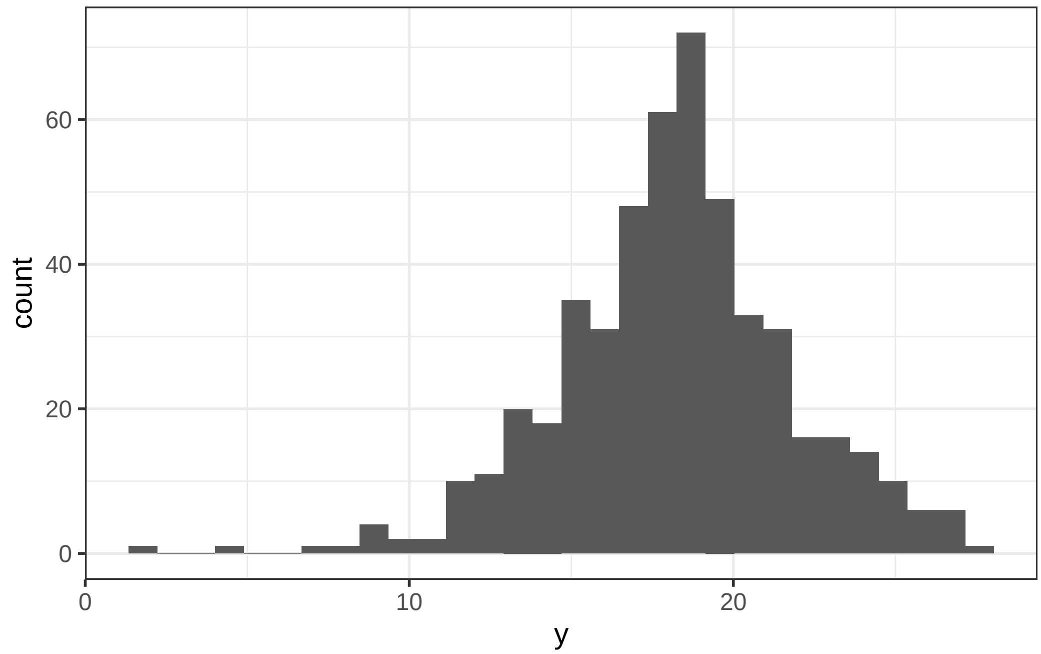
Linear regression

Consider the data set `reg_data_1.csv`. It contains the following variables:

- `y`: ice cream consumption in litres per year
- `x1`: Temperature in 10 degrees Celsius
- `x2`: Income in 1000 EUR
- `x3`: Height in cm

Study how ice cream consumption is associated with the explanatory variables and derive a sensive lineare regression model. Briefly justify your model specification.

```
dist_y <- ggplot(data = reg_data, mapping = aes(x=y)) +
  geom_histogram() + theme_bw()
dist_y
```

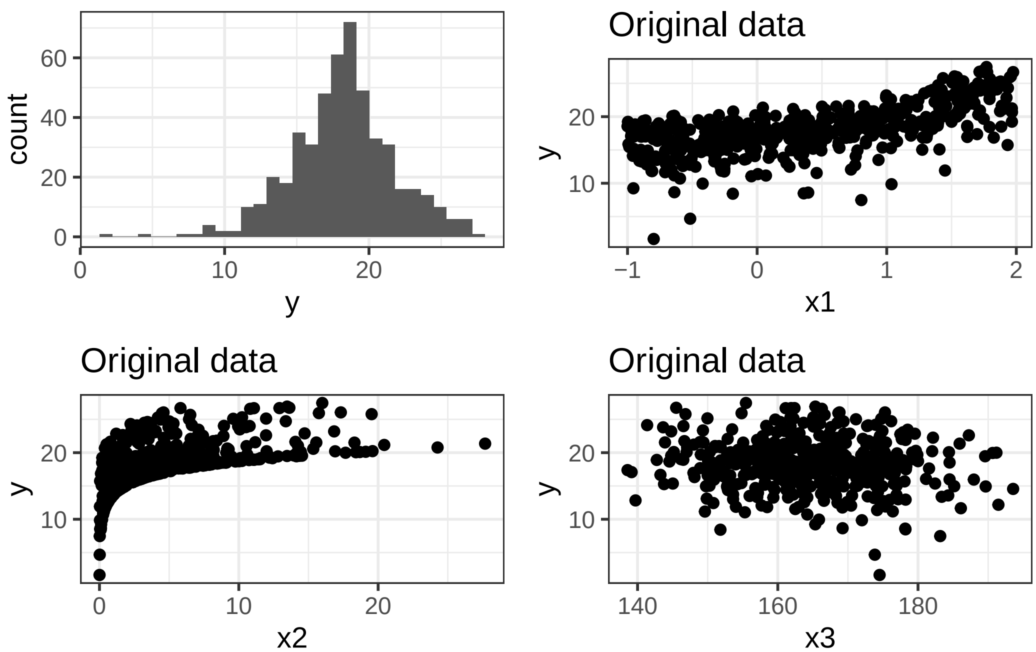



```r
linlin_plot_x1 <- ggplot(data = reg_data, mapping = aes(x=x1, y=y)) +
  geom_point() +
  labs(title = "Original data") +
  theme_bw()

linlin_plot_x2 <- ggplot(data = reg_data, mapping = aes(x=x2, y=y)) +
  geom_point() +
  labs(title = "Original data") +
  theme_bw()

linlin_plot_x3 <- ggplot(data = reg_data, mapping = aes(x=x3, y=y)) +
  geom_point() +
  labs(title = "Original data") +
  theme_bw()


ggarrange(
  dist_y, linlin_plot_x1, linlin_plot_x2,
  linlin_plot_x3, ncol = 2, nrow = 2)
```




```r
lm_correct <- lm(y~x1+I(x1**2)+log(x2), data = reg_data)
summary(lm_correct)
```

```
Call:
lm(formula = y ~ x1 + I(x1^2) + log(x2), data = reg_data)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.032602 -0.007347 -0.000002  0.007745  0.026055 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.400e+01  7.435e-04   18832   <2e-16 ***
x1          1.499e+00  8.354e-04    1795   <2e-16 ***
I(x1^2)     1.500e+00  6.967e-04    2153   <2e-16 ***
log(x2)     2.199e+00  3.650e-04    6026   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01008 on 496 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1 
F-statistic: 2.121e+07 on 3 and 496 DF,  p-value: < 2.2e-16
```

```
TA_correct <- ggplot(
  tibble("Fitted"=lm_correct$fitted.values,
         "Residuals"=lm_correct$residuals),
  mapping = aes(x=Fitted, y=Residuals)
  ) +
  labs(title = "TA plot for correct specification") +
  geom_point() +
  theme_bw()
TA_correct
```