ADA2: Class 13, Ch 08, polynomial regression

Name Here

March 04, 2022

Hooker's Himalayian boiling point altitude data

Dr. Joseph Hooker collected the following data in the 1840s on the boiling point of water and the atmospheric pressure at 31 locations in the Himalayas. Boiling point is measured in degrees Fahrenheit. The pressure is recorded in inches of mercury, adjusted for the difference between the ambient air temperature when he took the measurements and a standard temperature.

The goal was to develop a model to predict the atmospheric pressure from the boiling point.

Historical note: Hooker really wanted to estimate altitude above sea level from measurements of the boiling point of water. He knew that the altitude could be determined from the atmospheric pressure, measured with a barometer, with lower pressures corresponding to higher altitudes. His interest in the above modelling problem was motivated by the difficulty of transporting the fragile barometers of the 1840s. Measuring the boiling point would give travelers a quick way to estimate elevation, using the known relationship between elevation and barometric pressure, and the above model relating pressure to boiling point.

```r
library(erikmisc)
library(tidyverse)

dat_boil <-
  read_csv(
    "~/Dropbox/3_Education/Courses/stat_528_ada2/ADA2_CL_13_boilingpressure.csv"
  , skip = 2
  ) %>%
  mutate(
    boilingF_cen = boilingF - mean(boilingF)
  )

# x-variable mean for centering
dat_boil$boilingF %>% mean()
```

```
[1] 191.7871
```

```r
str(dat_boil)
```

```
spec_tbl_df [31 x 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ boilingF    : num [1:31] 211 210 208 202 201 ...
 $ pressure    : num [1:31] 29.2 28.6 28 24.7 23.7 ...
 $ boilingF_cen: num [1:31] 19.01 18.41 16.61 10.71 8.81 ...
 - attr(*, "spec")=
  .. cols(
  ..   boilingF = col_double(),
  ..   pressure = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

(2 p) Plot the data.

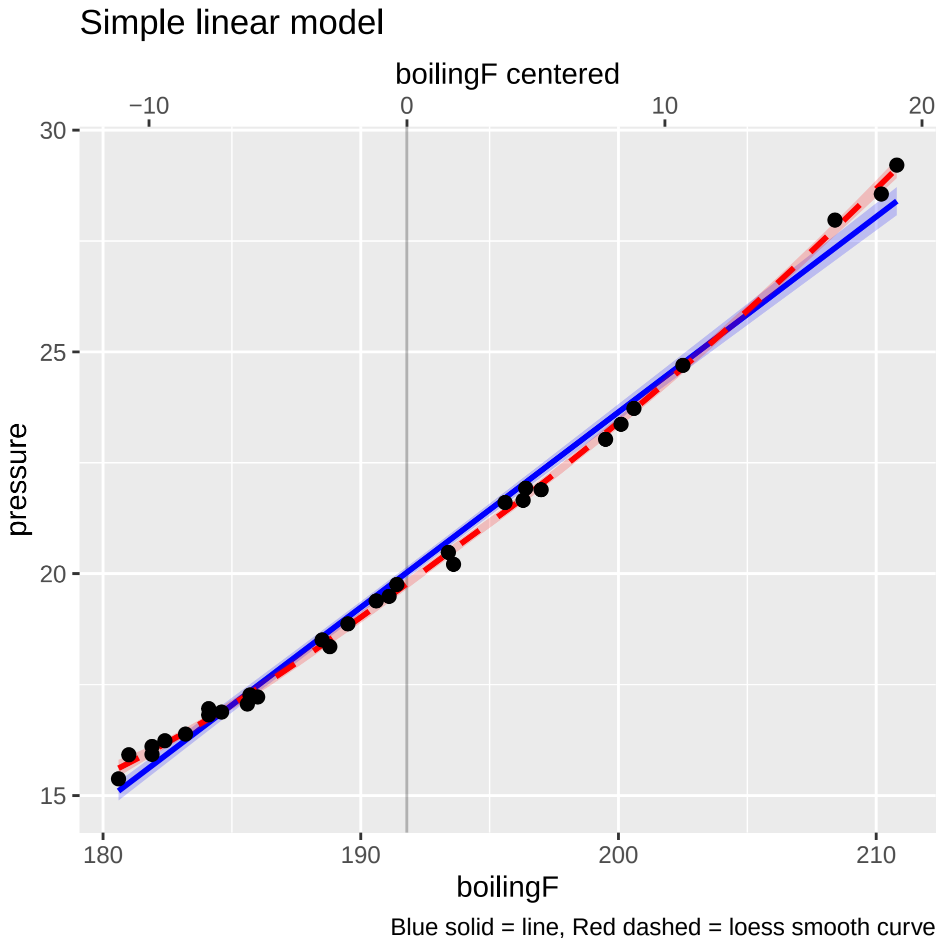
Using `ggplot`, try to implement these features in a plot. Overlay both a straight-line regression line in blue (`geom_smooth(method = lm, col = "blue", ...)`), as well as a loess smooth (default) dashed line in red (`geom_smooth(method = loess, col = "red", linetype = 2, ...)`). Using `alpha=1/5` will make the confidence bands more transparent. Also, if you plot the points last, they'll lie on top of the lines.

Describe the key features of this plot.

Solution

I'll give you this first plot to help get started, in particular to illustrate a nice use of the caption and the annotation of a second x-axis for the centered version of the `boilingF` variable.

```
library(ggplot2)
p <- ggplot(dat_boil, aes(x = boilingF, y = pressure))
p <- p + scale_x_continuous(sec.axis = sec_axis(~ . - mean(dat_boil$boilingF), name = "boilingF c
p <- p + geom_vline(xintercept = mean(dat_boil$boilingF), alpha = 1/4)
p <- p + geom_smooth(method = lm, se = TRUE, col = "blue", fill = "blue", alpha = 1/5)
p <- p + geom_smooth(method = loess, se = TRUE, col = "red", fill = "red", linetype = 2, alpha =
p <- p + geom_point(size = 2)
p <- p + labs(title = "Simple linear model"
             , caption = "Blue solid = line, Red dashed = loess smooth curve"
               )
print(p)
```



Well, these are very straight lines. However, you can see that the LOESS smoother demonstrates some quadratic behavior, where a linear fit underpredicts at the boundaries, and overpredicts toward the center. Maybe there's something to it. Let's see.

(3 p) Fit a simple linear regression, assess assumptions.

Fit a simple linear regression model for predicting pressure from boiling point. Provide output for examining residuals, outliers, and influential cases.

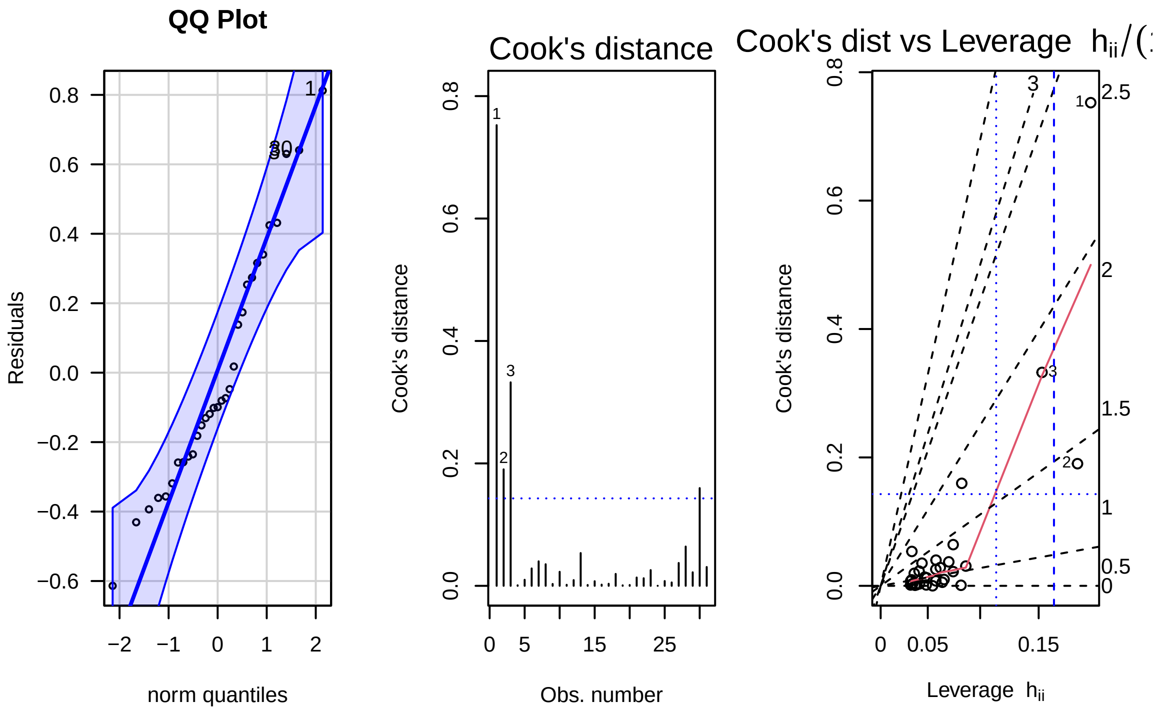
Looking at the plots, are there any indications that the mean pressure is not linearly related to boiling point? Are there any observations that appear to be highly influencing the fit of this model? Are there certain points or regions of the data where the model does not appear to fit well? Discuss.

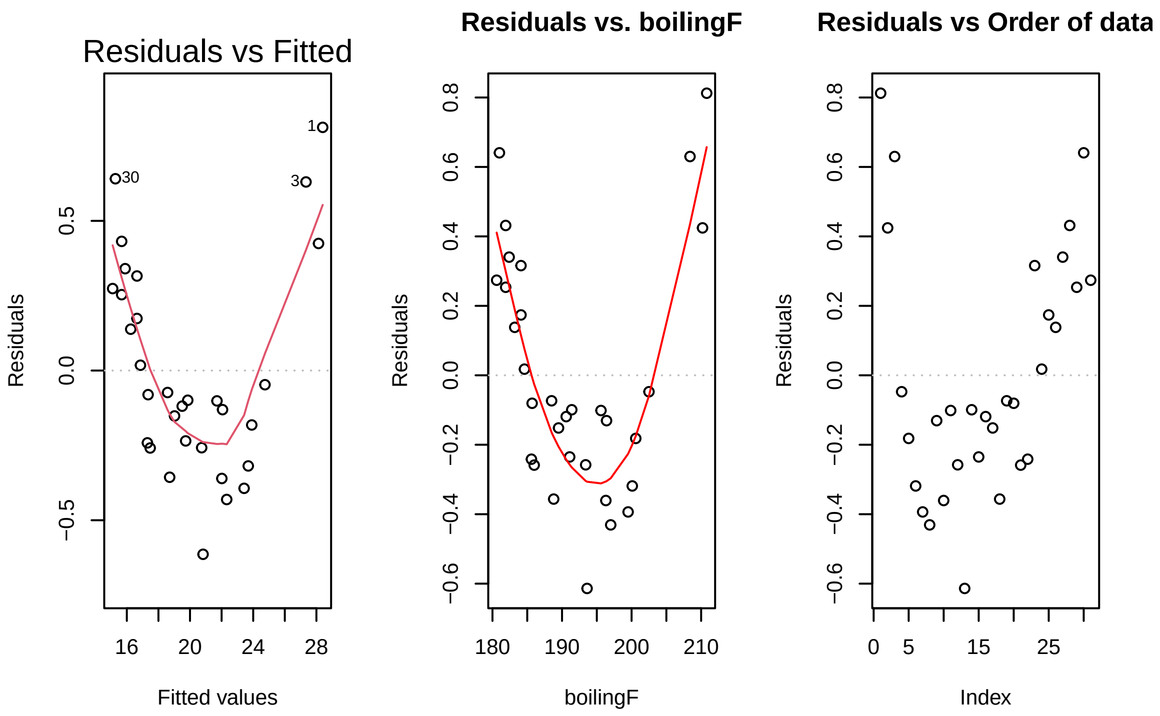
Which, if any, of the standard linear regression model assumptions appears to be violated in this analysis? If you believe that some of the assumptions are violated, does it appear that deleting one or two points would dramatically improve the fit? Would you use this model for predicting pressure from boiling point? Discuss and carry out any needed analysis to support your position.

Solution

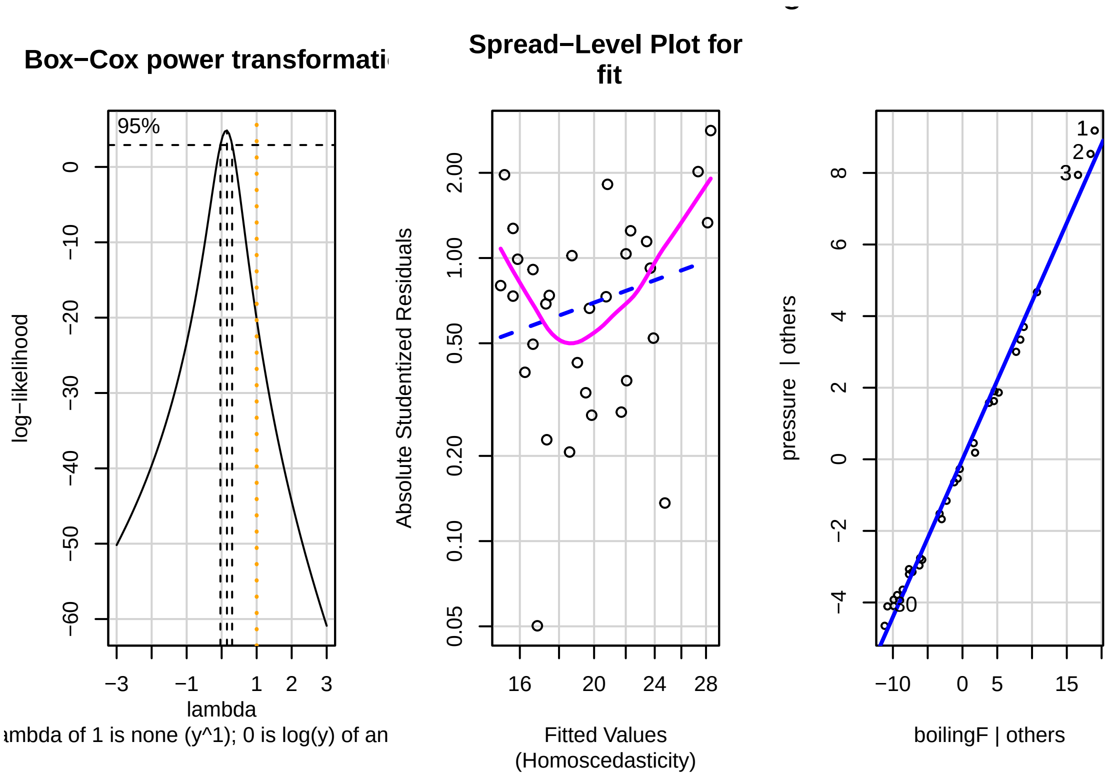
```
mod1 <- lm(pressure ~ boilingF, data = dat_boil)

e_plot_lm_diagostics(mod1)
```





```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.94974, Df = 1, p = 0.046879
```



Oh yeah, those are some very alarming diagnostics. As I suggested above, the residuals are showing a U-shaped pattern across fitted values, which is clearly driven by the pattern across boiling temperature. The model is over-predicting toward the boundaries and under-predicting toward the center of boiling temperature.

There are also some severe outliers, but let's try to fix the fit first and see what happens. It could solve all our problems.

(1 p) Interpret R^2

Interpret R^2 in the previous simple linear regression model.

Solution

```
summary(mod1)
```

```

Call:
lm(formula = pressure ~ boilingF, data = dat_boil)

Residuals:
     Min       1Q   Median       3Q      Max
-0.61383 -0.24968 -0.09921  0.26365  0.81232

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.412751   1.429165  -45.07   <2e-16 ***
boilingF      0.440282   0.007444   59.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3563 on 29 degrees of freedom
Multiple R-squared:  0.9918,    Adjusted R-squared:  0.9915
F-statistic:  3498 on 1 and 29 DF,  p-value: < 2.2e-16
```

Great example. The R^2 shows the linear relationship between temperature and pressure explains over 99% of the variation in pressure. So although we have evidence that we can do better, this model would probably do rather well predicting pressure from boiling temperature.

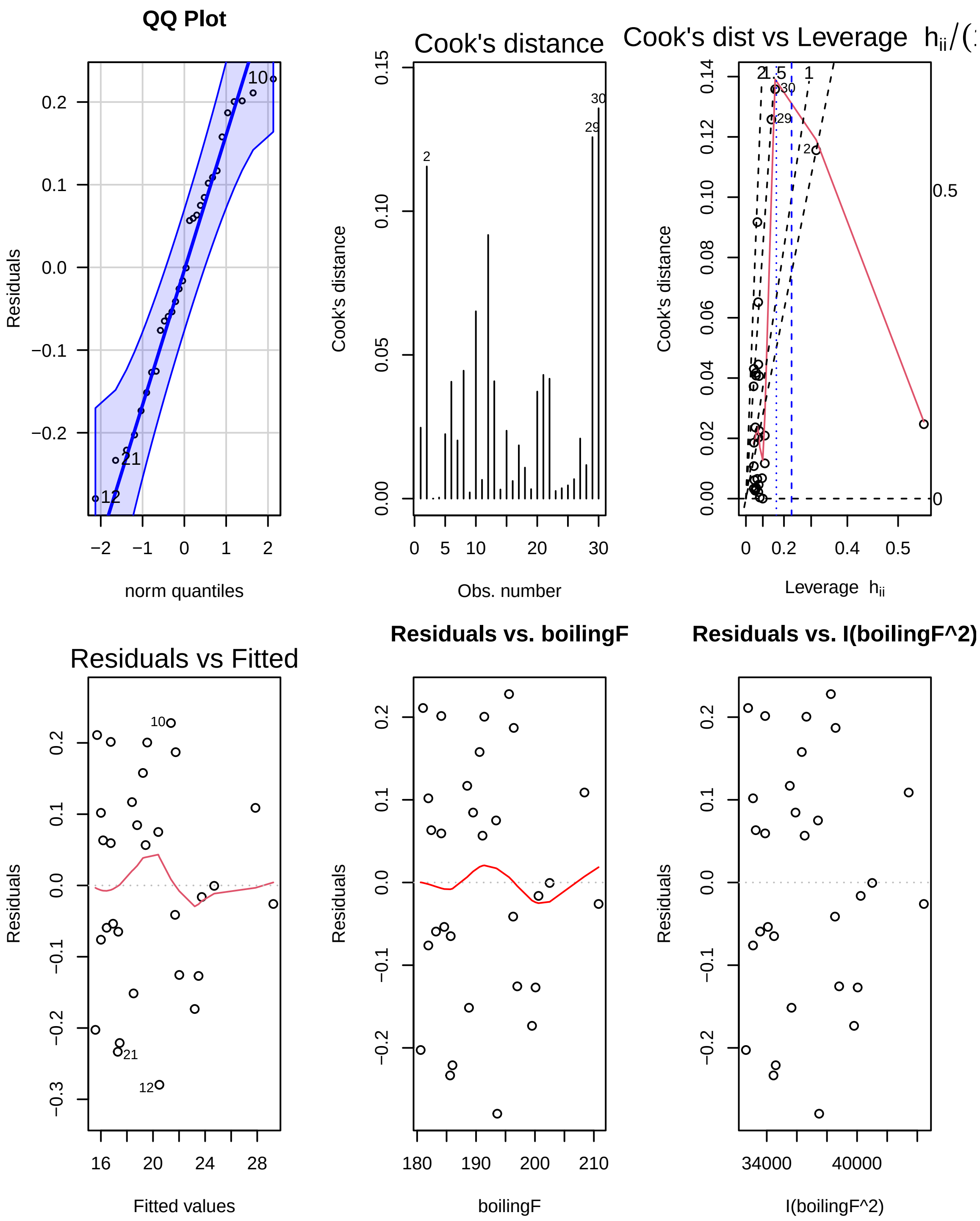
(2 p) A better model.

Decide whether transformation, or a polynomial model in boiling point, is needed to adequately summarize the relationship between pressure and boiling point. If so, perform a complete analysis of the data on this scale (that is, check for influential observations, outliers, non-normality, etc.).

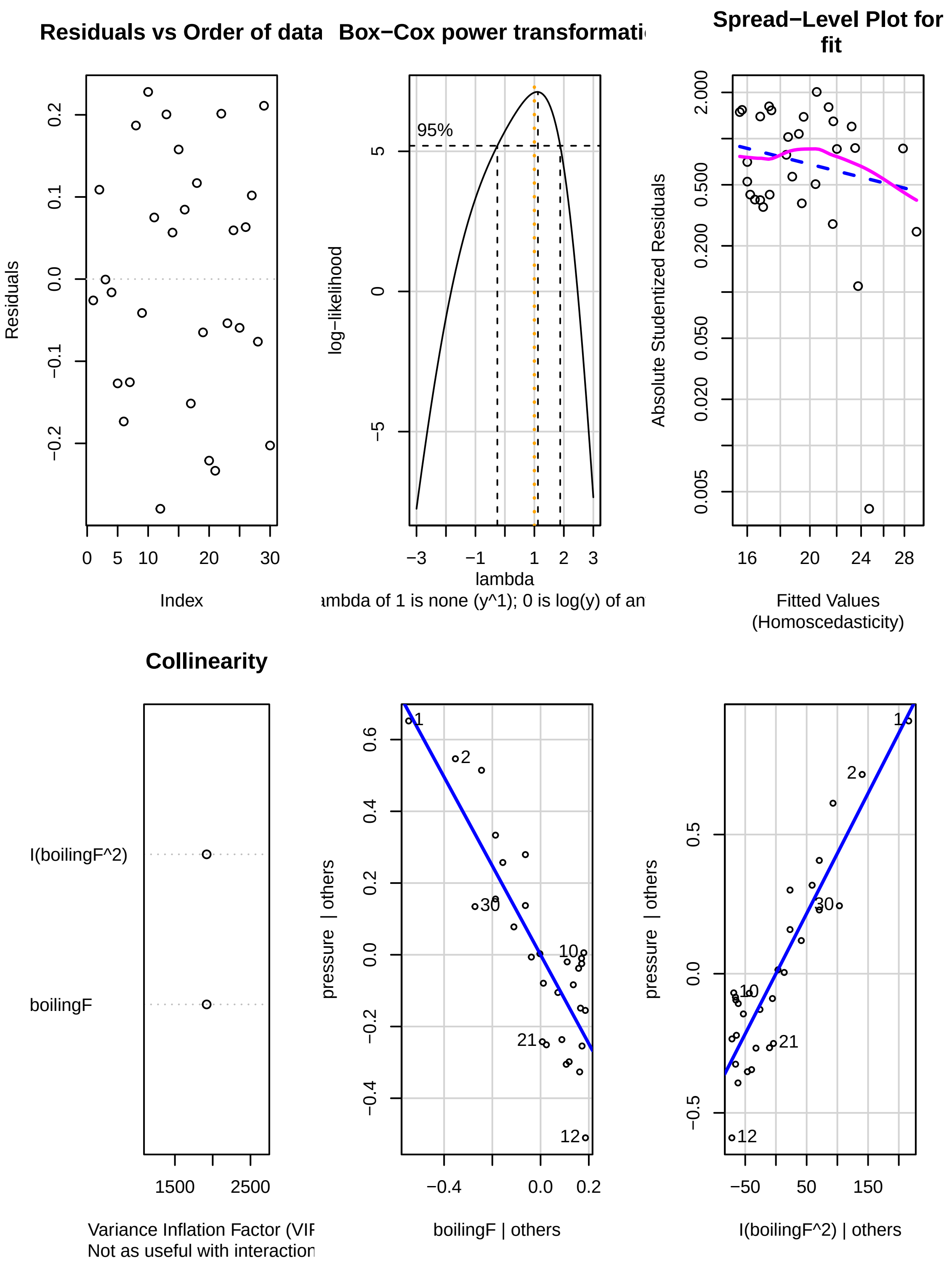
Solution

```
dat_boil <- dat_boil %>% slice(-2)
mod2 <- lm(pressure ~ boilingF + I(boilingF^2), data = dat_boil)

e_plot_lm_diagostics(mod2)
```



```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.4581783, Df = 1, p = 0.49848
```



```
summary(mod2)
```

```
Call:
lm(formula = pressure ~ boilingF + I(boilingF^2), data = dat_boil)

Residuals:
      Min        1Q    Median        3Q       Max
-0.279633 -0.113221 -0.008379  0.107087  0.227856

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   98.4383068 14.6232453   6.732 3.16e-07 ***
boilingF      -1.2393458  0.1510596  -8.204 8.25e-09 ***
```

```
I(boilingF^2)  0.0043219  0.0003896  11.094 1.46e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1519 on 27 degrees of freedom
Multiple R-squared:  0.9983,    Adjusted R-squared:  0.9982
F-statistic:  8062 on 2 and 27 DF,  p-value: < 2.2e-16
```

I decided to fit a quadratic for temperature, and that solved a lot of our problems. There was a rather extreme value that may have been an outlier, so I've removed it, but the model didn't change much. We now see no patterns whatsover i the residuals, normally distributed error, and no clear outliers.

The quadratic fit is extremely significant at $\alpha = 0.05$, and our $adj - R^2$ value has increased from 0.991 to 0.998. Not a huge improvement with respect to predictive ability, but we ought to think hard about the theory behind all this. Did we anticipate a quadratic fit? Does it make sense? Need we report our results to the theoretical physicists?

(2 p) Final model.

Regardless of which scale you choose for the analysis, provide an equation to predict pressure from boiling point. Write a short summary, pointing out any limitations of your analysis.

Solution

$$\widehat{pressure} = 98.44 - 1.24 \times TEMP + 0.0043 \times TEMP^2$$

I've already said a lot. Personally, I think limitations might be more about confounding factors in the experimental design. This work was hard to do, and I'm unconvinced that the quadratic fit isn't some artifact of how the experiment was carried out. Otherwise ... I feel pretty good about this one!

Example based on the first linear model *Assuming you called your linear model object `lm_p_b1`, then the equation with code below will place the intercept and slope in the equation. Just add an `r` before each of the `signif(...)` inline code chunks to make the numbers appear. Then use this example to write your final model here.*

THIS CODE IS BROKEN FOR SOME REASON. REMOVING.