ADA2: Class 04, Ch 02 Introduction to Multiple Linear Regression

Tim Farkas

January 31, 2022

Water Usage of Production Plant

A production plant cost-control engineer is responsible for cost reduction. One of the costly items in his plant is the amount of water used by the production facilities each month. He decided to investigate water usage by collecting seventeen observations on his plant's water usage and other variables.

Variable	Description
Temperature	Average monthly temperate (F)
Production	Amount of production (M pounds)
Days	Number of plant operating days in the month
Persons	Number of persons on the monthly plant payroll
Water	Monthly water usage (gallons)

```
library(erikmisc)
library(tidyverse)

# First, download the data to your computer,
#   save in the same folder as this Rmd file.

# read the data
dat_water <- read_csv("~/Dropbox/3_Education/Courses/stat_528_ada2/ADA2_CL_04_water.csv")
str(dat_water)
```

```
spec_tbl_df [17 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Temperature: num [1:17] 58.8 65.2 70.9 77.4 79.3 81 71.9 63.9 54.5 39.5 ...
 $ Production : num [1:17] 7107 6373 6796 9208 14792 ...
 $ Days       : num [1:17] 21 22 22 20 25 23 20 23 20 20 ...
 $ Persons    : num [1:17] 129 141 153 166 193 189 175 186 190 187 ...
 $ Water      : num [1:17] 3067 2828 2891 2994 3082 ...
 - attr(*, "spec")=
  .. cols(
  ..   Temperature = col_double(),
  ..   Production = col_double(),
  ..   Days = col_double(),
  ..   Persons = col_double(),
  ..   Water = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

```
#dat_water

dat_water <-
  dat_water %>%
  mutate(
    # Add an ID column
    id = 1:n(), .before=1
  ) %>%
  # filter to remove observations (if needed)
  #   TRUE        indicates that ALL observations are included
  #   !(id %in% c(4, 12)) indicates to include all observations that are NOT id 4 or 12
  filter(
    TRUE  # !(id %in% c(4, 12))
  )
```

Note: Because of the high correlation between `Production` and `Persons`, do not include `Persons` in the model.

Rubric

Following the in-class assignment this week, perform a complete multiple regression analysis.

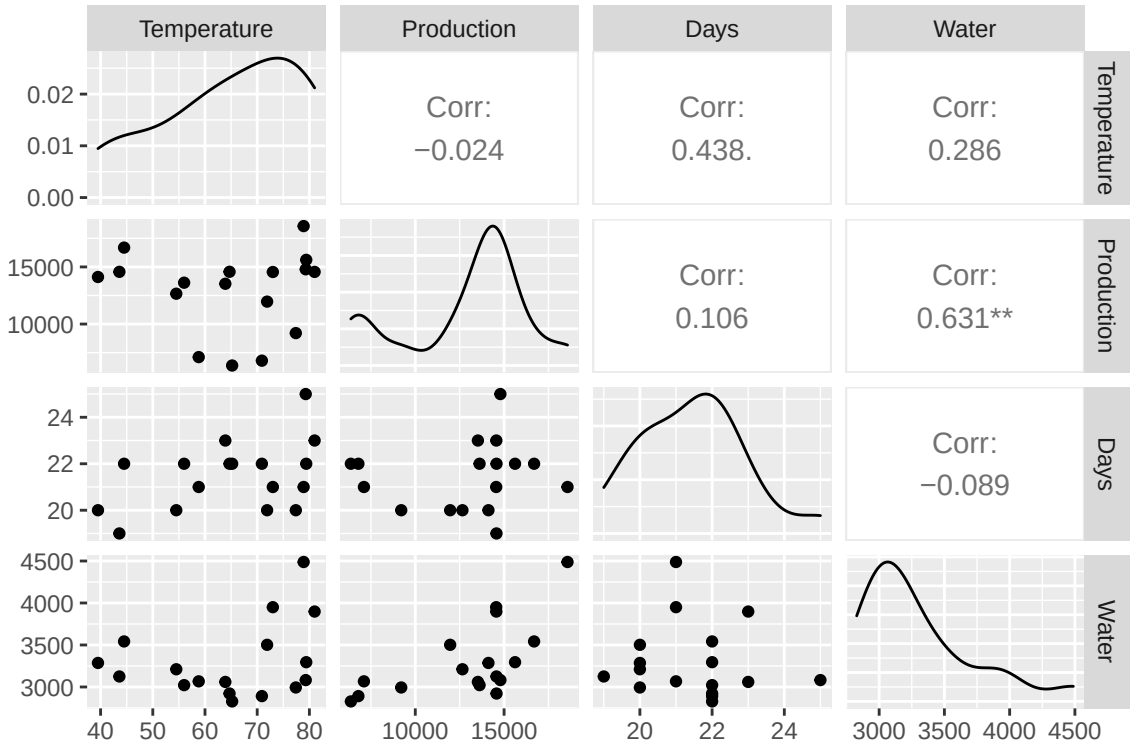
1. (1 p) Scatterplot matrix and interpretation
2. (2 p) Fit model, assess multiple regression assumptions
3. (1 p) Interpret added variable plots
4. (1 p) If there are model assumption issues, say how you address them at the beginning and start again.
5. (1 p) State and interpret the multiple regression hypothesis tests
6. (2 p) Interpret the significant multiple regression coefficients
7. (1 p) Interpret the multiple regression R^2
8. (1 p) One- or two-sentence summary

Solutions

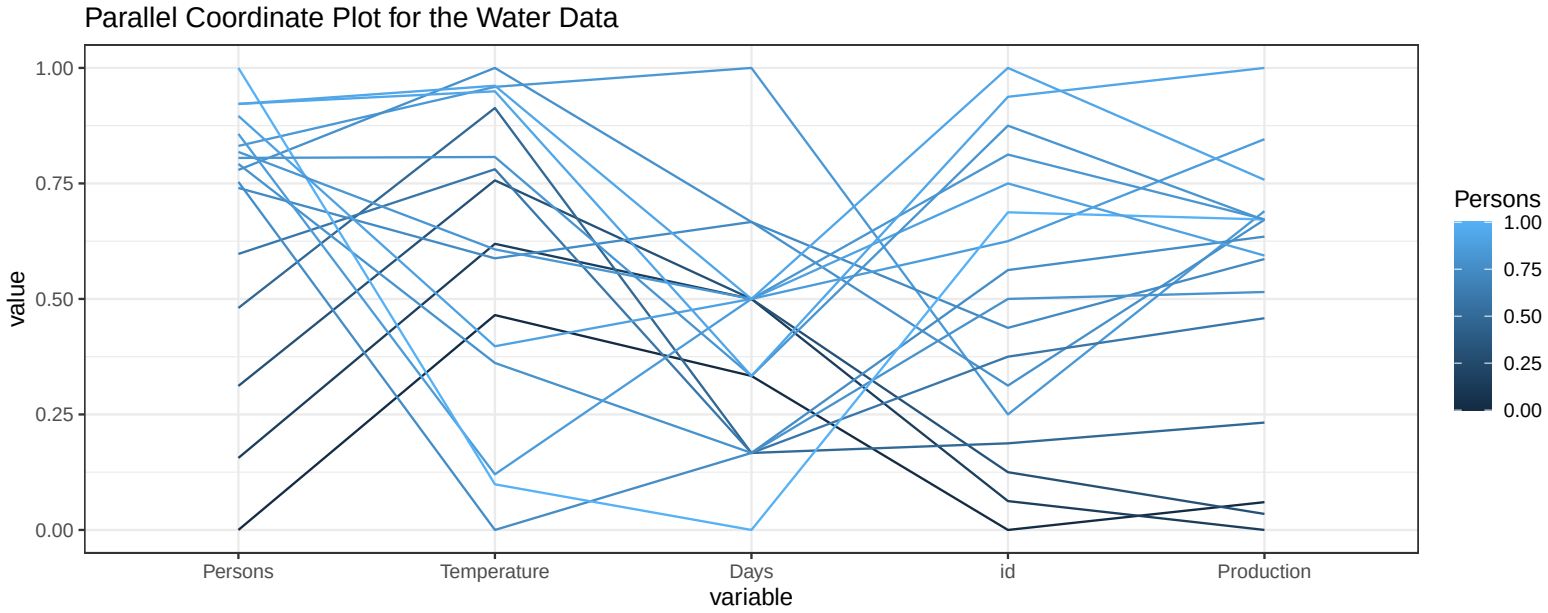
(1 p) Scatterplot matrix

In a scatterplot matrix below interpret the relationship between each pair of variables. If a transformation is suggested by the plot (that is, because there is a curved relationship), also plot the data on the transformed scale and perform the following analysis on the transformed scale. Otherwise indicate that no transformation is necessary.

```
library(ggplot2)
library(GGally)
#p <- ggpairs(dat_water)
p <- ggpairs(dat_water %>% select(-id, -Persons))  ## use select to remove vars
print(p)
```



A parallel coordinate plot is another way of seeing patterns of observations over a range of variables.

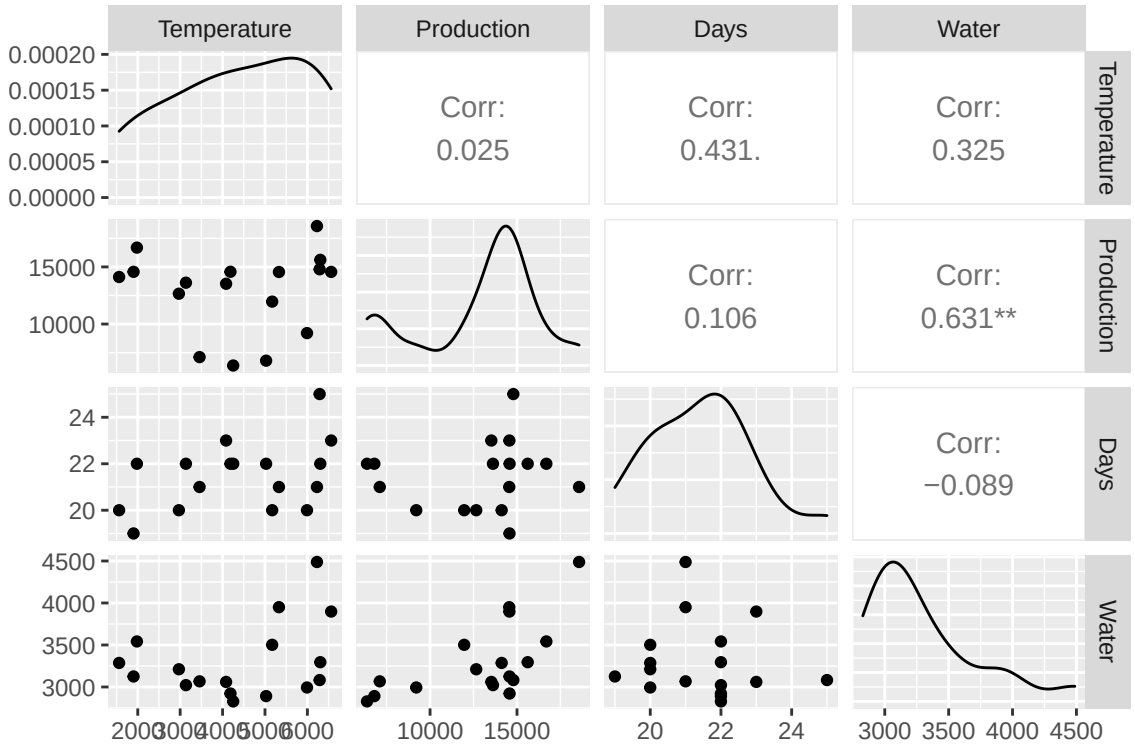


Solution

There appears to be a straight-forward, positive correlation between production and water use, which makes good sense. There might be a positive relationship between temperature and water use, which also makes sense, but the relationship looks nonlinear, so below we transform with the square of temperature and reexamine the correlations. There's a hint of positive correlation between temperature and days, but that doesn't make much sense, and the relationship looks weak. No other correlations are apparent based on this pairwise plot.

```
p <- dat_water %>%
  select(-c(id, Persons)) %>% #View
  mutate(across(Temperature, ~ .x^2)) %>%
  ggpairs()

print(p)
```



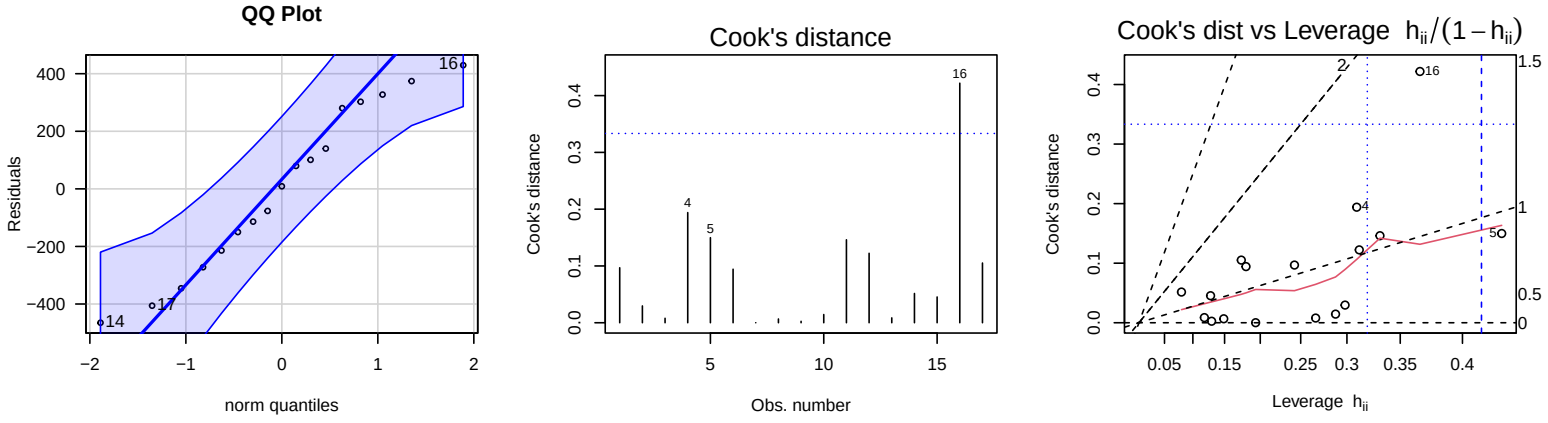
Umm ... this square transformation didn't seem to address the non-linearity, or do very much at all. I also tried an exponential transformation, which was a total disaster.

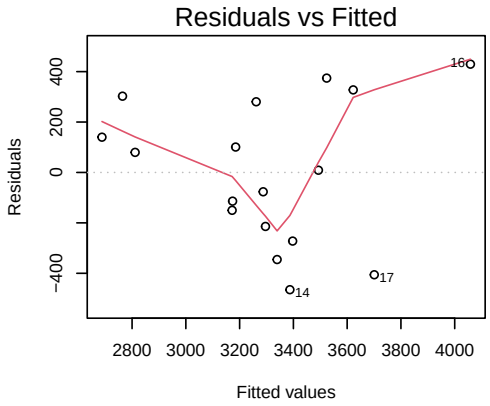
(2 p) Multiple regression assumptions (assessing model fit)

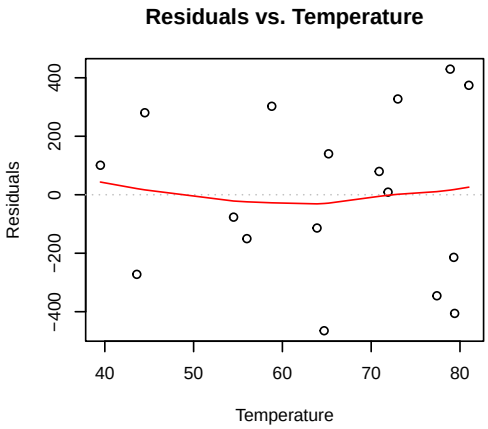
Below the multiple regression is fit. Start by assessing the model assumptions by interpretting what you learn from the first seven plots (save the added variable plots for the next question). If assumptions are not met, attempt to address by transforming a variable (or removing an outlier) and restart at the beginning using the new transformed variable.

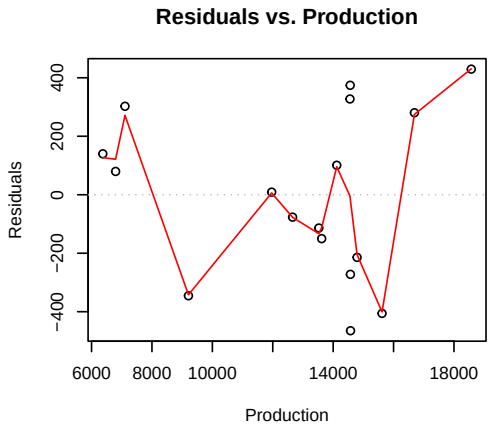
```
# fit the simple linear regression model
#lm_w_tpdp <- lm(Water ~ Temperature + Production + Days + Persons, data = dat_water)
lm_w_tpd <-
  lm(
    Water ~ Temperature + Production + Days
  , data = dat_water
  )
```

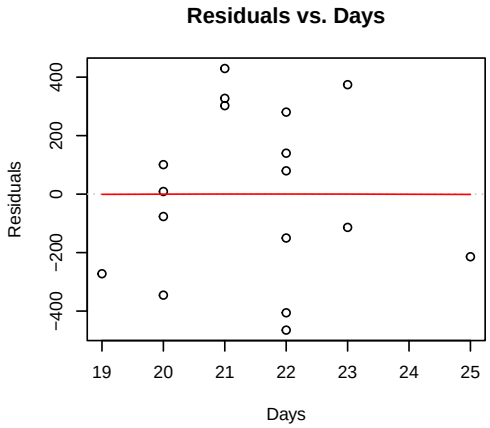
Plot diagnostics.

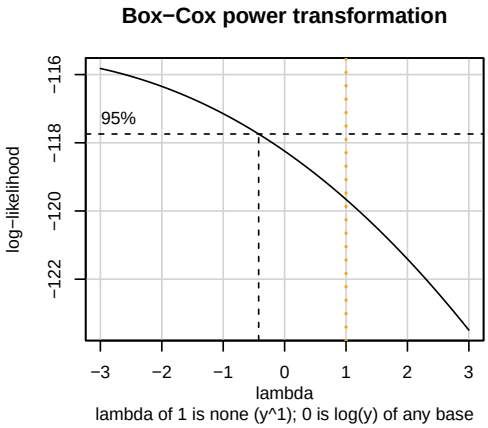


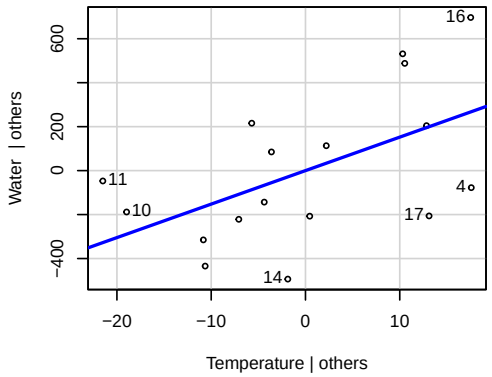


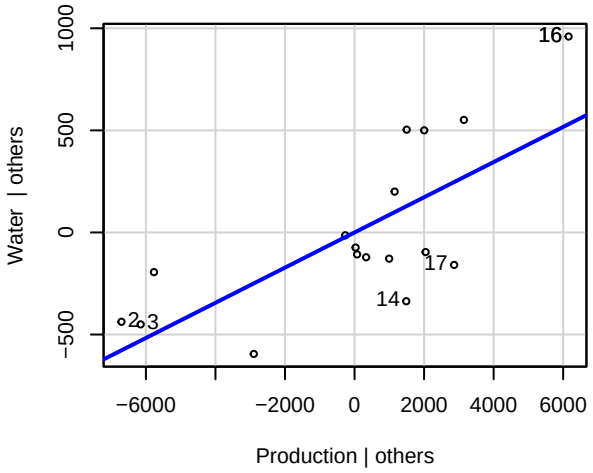


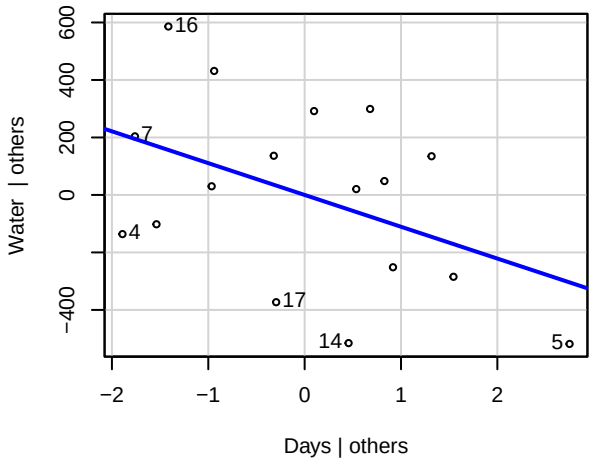












Solution

[answer]

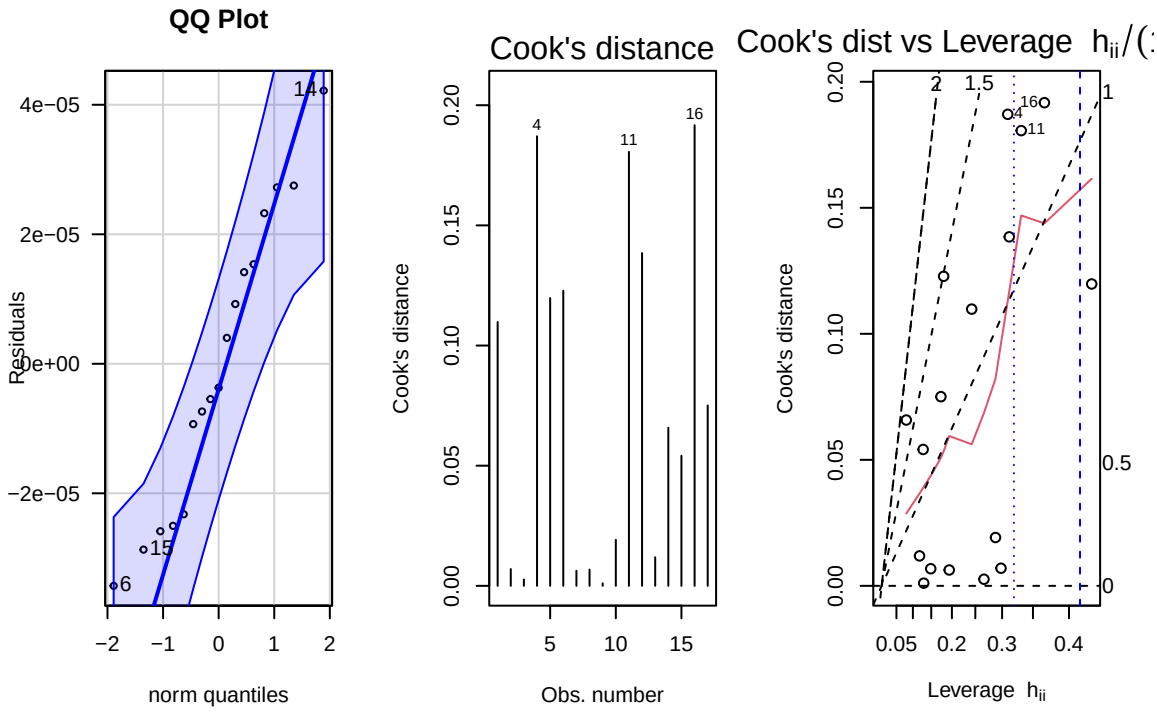
From the diagnostic plots above,

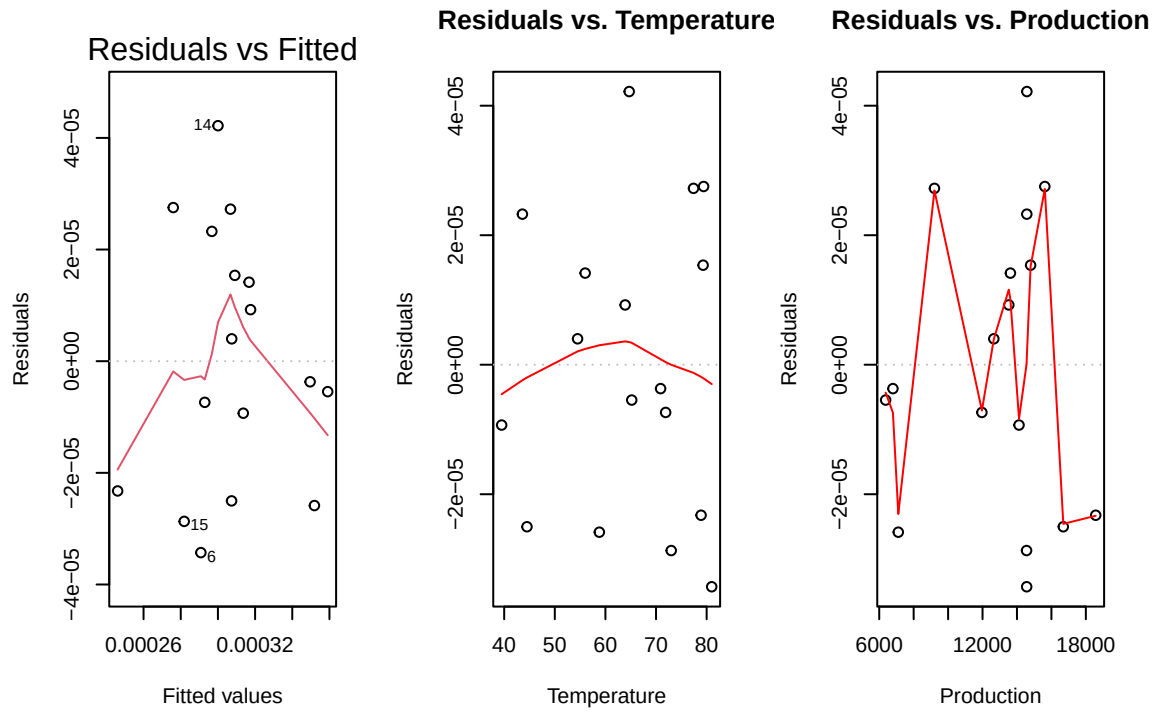
(1) The results actually look normal enough for me.
(2) There is definitely one major outlier based on Cook's distance,
(3) also visible here, though other points with high leverage don't influence the overall fit of the model.
(4) I think we're seeing here unstable variace, showing a bit of a fan.
(5) Looks fine.
(6) Also looks okay, I think, but this red line is all wonky.
(7) Fine.

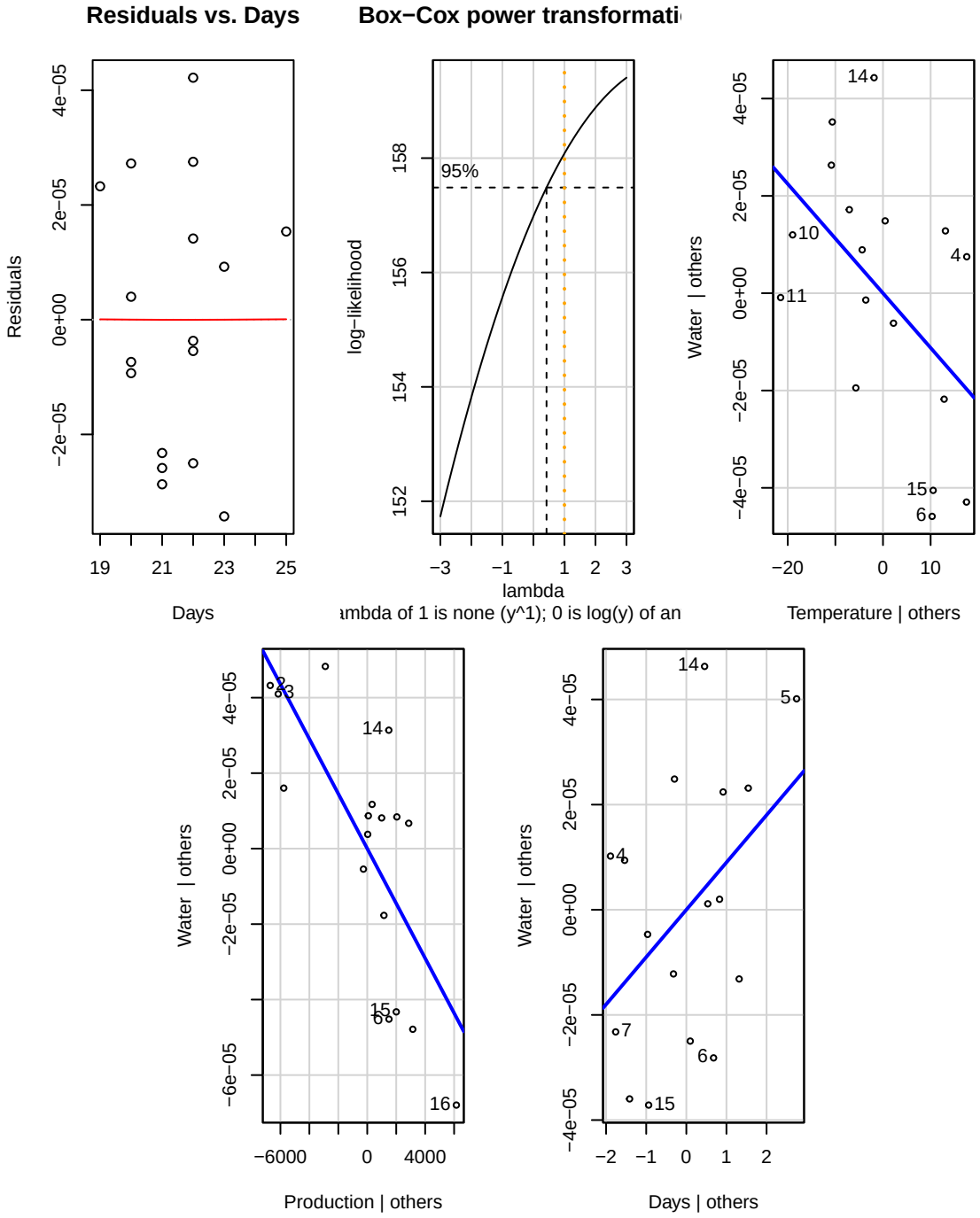
Let's try a tranformation following the BoxCox profile:

```r
dat_water_t <- dat_water %>%
  mutate(across(Water, ~.x^-1))

lmt <-lm(Water ~ Temperature + Production + Days, data = dat_water_t)
  e_plot_lm_diagostics(lmt, sw_plot_set = "simpleAV")
```







OK, that seems to have helped! We don't really have much fanning of the residuals, and the outlier is now not so extreme. Just remember the effects will be switched, because we're working with the reciprocal of water use now.

(1 p) Added variable plots

Use partial regression residual plots (added variable plots) to check for the need for transformations. If linearity is not supported, address and restart at the beginning.

Solution

The AV plots show a negative effect of temperature and production on the reciprocal of water use, and a positive effect of days. Not so sure about this effect of day ... doesn't make much sense to me. All effects are rather linear.

(1 p) Multiple regression hypothesis tests

State the hypothesis test and conclusion for each regression coefficient.

```
# use summary() to get t-tests of parameters (slope, intercept)
summary(lmt)
```

```
Call:
lm(formula = Water ~ Temperature + Production + Days, data = dat_water_t)

Residuals:
       Min         1Q     Median         3Q        Max
-3.429e-05 -2.324e-05 -3.691e-06  1.537e-05  4.219e-05

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.805e-04  9.489e-05   2.956  0.01115 *
Temperature -1.123e-06  5.240e-07  -2.142  0.05169 .
Production  -7.264e-09  1.815e-09  -4.002  0.00151 **
Days         9.004e-06  4.866e-06   1.851  0.08709 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.538e-05 on 13 degrees of freedom
Multiple R-squared:  0.6029,    Adjusted R-squared:  0.5113
F-statistic: 6.579 on 3 and 13 DF,  p-value: 0.006077
```

Solution

Ignoring the intercept:

The hypotheses for each coefficient are for whether the coefficient values are different from zero, conditional on all other predictors. In the case of the intercept, that's the value of water use when all other variables are at 0 (not interesting). For the other three, it's essentially whether there is an effect of the variable on water use. We see a significant negative effect of production on the reciprocal of water use, and marginal effects of temperature and days, in the negative and positive direction respectively. Marginal here defined as rejecting the null hypothesis at $\alpha = 0.10$, but failing to do so at $\alpha = 0.05$.

For $H_0 : \beta_{\text{Temperature}} = 0$, the t-statistic is -2.142 with an associated p-value of 0.05169. Thus, we fail to reject H_0 concluding that the slope is statistically significantly different from 0 conditional on the other variables in the model.

Similarly, for $H_0 : \beta_{\text{Production}} = 0$, the t-statistic is -4.002 with an associated p-value of 0.001507. Thus, we reject H_0 concluding that the slope is statistically significantly different from 0 conditional on the other variables in the model.

Similarly, for $H_0 : \beta_{\text{Days}} = 0$, the t-statistic is 1.851 with an associated p-value of 0.08709. Thus, we fail to reject H_0 concluding that the slope is statistically significantly different from 0 conditional on the other variables in the model.

(1 p) Multiple regression interpret coefficients

Interpret the significant coefficients of the multiple regression model.

Solution

If we only consider the effect of Production significant, the coefficient indicates that for every one unit increase in production, the reciprocal of water use increases by -7.264×10^{-9}.

(1 p) Multiple regression R^2

Interpret the Multiple R-squared value.

Solution

The multiple R^2 is 0.6029, indicating about 60% of variation in the inverse of water use is explained by this model.

(1 p) Summary

Summarize your findings in one sentence.

Solution

We have strong evidence that the amount of production increases water use, and some weaker evidence temperature increases water use, and that the number of days decreases water use.

Unused plots

```
## Aside: While I generally recommend against 3D plots for a variety of reasons,
## so you can visualize the surface fit in 3D, here's a 3D version of the plot.
## I will point out a feature in this plot that we would't see in other plots
## and would typically only be detected by careful consideration
## of a "more complicated" second-order model that includes curvature.

# library(rgl)
# library(car)
# scatter3d(Water ~ Temperature + Production, data = dat_water)
```

These bivariate plots can help show the relationships between the response and predictor variables and identify each observation.