Lab2: Linear Regression

MAT43 Statistical Machine Learning

Silvia Montagna

3/20/2018

The first part of this lab involves the UN data set from ALR. Download the `alr3` library and load the data to answer the following questions. Please add appropriate code to the chuncks to suppress messages and warnings as needed once you are sure the code is working properly and remove instructions if no longer needed.

```
library(alr3)
data(UN3)
help(UN3)
```

In the second part of this lab, you will explore more advanced `R` graphic tools such as *Added variable* and *term* plots. These are useful to gain insights on the importance of each predictor in the context of linear regression, and whether transformations of the response &/or the predictors are necessary.

Exploratory Data Analysis

1. Create a summary of the data. How many variables have missing data? Which are quantitative and which are qualitative?
2. What is the mean and standard deviation of each quantitative predictor? Provide your results in a nicely formatted table.
3. Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings regarding trying to predict `ModernC` from the other variables. Are there potential outliers, nonlinear relationships or transformations that appear to be needed?

Model fitting

4. Use the `lm()` function to perform a multiple linear regression with `ModernC` as the response and all other variables as the predictors, using the formula `ModernC ~ .`, where the `.` includes all remaining variables in the dataframe. Create diagnostic residual plots from the linear model object and comment on results regarding assumptions.
5. Using the `powerTransform` function from the `car` library or graphical methods find appropriate transformations of the predictor variables to be used as predictors in the linear model. If any predictors are negative, you may need to transform them so that they are non-negative. Describe your method and the resulting transformations.
6. Given the selected transformations of the predictors, select a transformation of the response and justify.

7. Fit the regression using the transformed variables. Provide residual plots and comment. Provide summaries of coefficients with 95% confidence intervals in a nice table with interpretations.
8. Are there any outliers in the data? Explain. If so, refit the model after removing any outliers. *Hint: explore the* `outlierTest` *function*

Summary of results

9. Provide a brief paragraph summarizing your final model and findings suitable for the US envoy to the UN after adjusting for outliers or influential points.

Extra: Added Variable & Term plots with transformations

Added variable plots

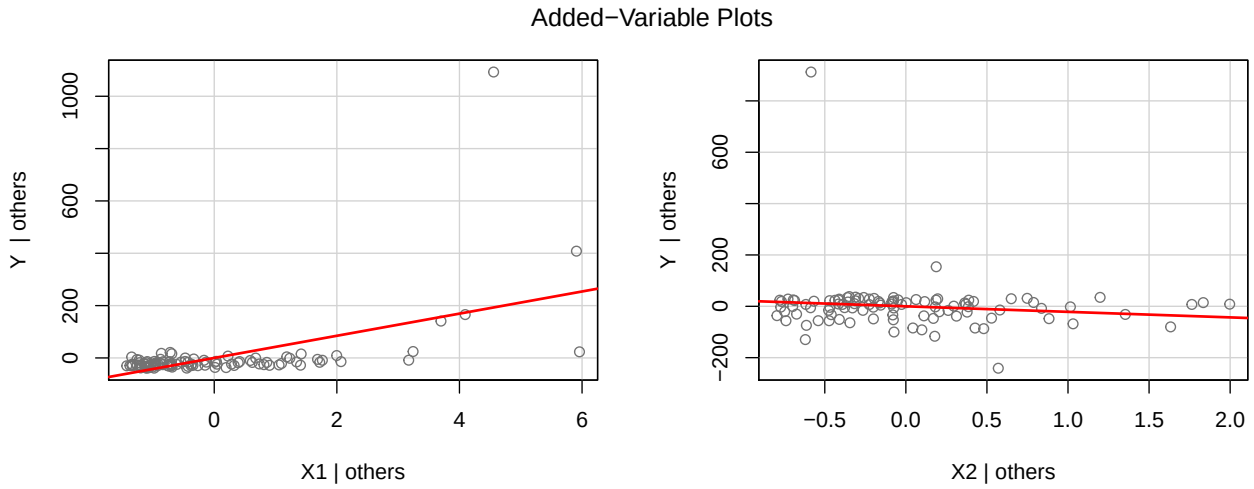
Added variable plots show the effect of adding another predictor to a linear model already having one or more independent variables. Added variable plots are formed by:

1. Computing the residuals of regressing the response variable against the independent variables but omitting Xj
2. Computing the residuals from regressing X_j against the remaining independent variables
3. Plotting the residuals from (1) against the residuals from (2)

Let's consider the following simulated example:

```
# Simulate some synthetic data
n = 100
logx1 = rnorm(n)
x1 = exp(logx1)
x2 = abs(rnorm(n))
logy = .5 + 3*logx1 - 2*sqrt(x2) + rnorm(n)
simdat = data.frame(X1=x1, X2 = x2, Y=exp(logy))

# From library(car)
library(car)
mod1 = lm(Y ~ X1 + X2, data = simdat)
avPlots(mod1, pch = 1, col = "grey45")
```



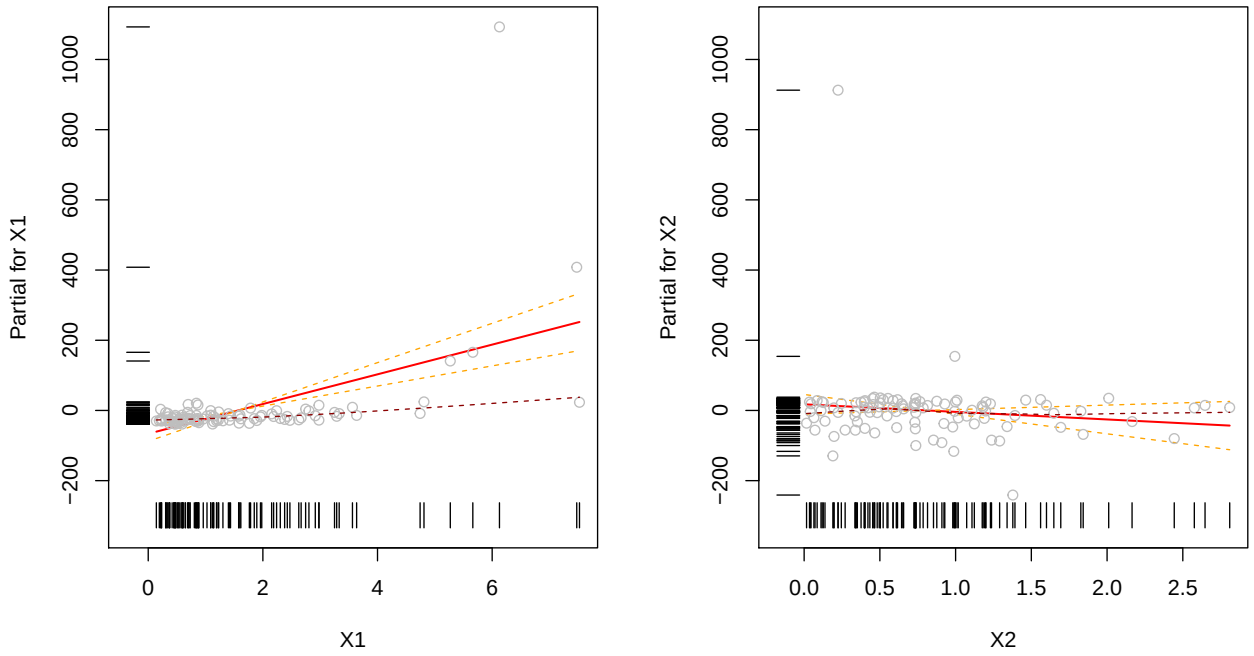
In the added variable plot on the left, we 1) regress Y on X_2 and find the residuals (y axis); 2) regress X_1 on X_2 and find those residuals (x axis); 3) plot the residuals of (1) against the residuals from (2). These are the points you see in the left panel. If X_1 can explain any residual variation after removing the effect of X_2 on both Y and X_1, then we should see a relationship in the plot. The slope of the line in the added variable plot is the adjusted regression coefficient. This plot can also show if the linear term in X_1 is appropriate or perhaps there is a nonlinear relationship.

Term plots

We now investigate the `termplot` function in `R`. Let's do so by continuing with the simulated example presented above:

```
# Linear regression summary
# summary(mod1)

# Term plots
par(mfrow = c(1,2))
termplot(mod1, partial.resid = T, se = T, rug = T, smooth = panel.smooth)
```



What is in a term plot?

- The x-axis is the (untrasformed) variable in your dataframe (X_1, X_2)
- The red line is the "term" of that variable's contribution to $f(x)$ (see below)
- The y-axis shows the partial residuals for that term
- `partial.resid = T` adds the partial residuals to the plot
- `rug = T` shows the location of the data on the axes
- `se = T` adds the standard error of the term's contribution to $f(x)$
- `smooth = panel.smooth` adds "smoothed" means to plot (magenta)

Terms

To understand term plots, it helps to know that the centered model:

$$Y = \overline{Y} + \widehat{\beta}_1(X_1 - \overline{X_1}) + \widehat{\beta}_2(X_2 - \overline{X_2}) + e \tag{1}$$

is equivalent to:

$$Y = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + e \tag{2}$$

Equation (1) has $\overline{Y}$ as the intercept, and centered predictors. If you take the centered predictor values and multiply them by their respective coefficients you get what are referred to in `R` as *terms*:

$$\widehat{\beta}_1(X_1 - \overline{X_1})$$

$$\widehat{\beta}_2(X_2 - \overline{X_2})$$

The *partial residuals* (y-axis of a term plot) are the sum of the residuals (difference between the observed and fitted response values, e) and the predictor terms.

So in a term plot, `R` takes the terms for each predictor, adds the residuals to the terms to create partial residuals, and then plots partial residuals versus their respective predictor (if you specify `partial.resid = TRUE`). The fitted line is the result of regressing the predictor's partial residuals on the predictor itself. The slope of the line in the term plot graphs is the coefficient in the summary output.

Ultimately, a term plot splits the response value into different parts: an overall mean, a term that is due to X_1, a term that is due to X_2, and the residual. To create the partial residual for X_1, we add the X_1 *term* and the residual. *This sum accounts for the part of the response not explained by the other terms*:

$$Y - (\overline{Y} + \widehat{\beta}_2(X_2 - \overline{X}_2)) = \widehat{\beta}_1(X_1 - \overline{X}_1) + e$$

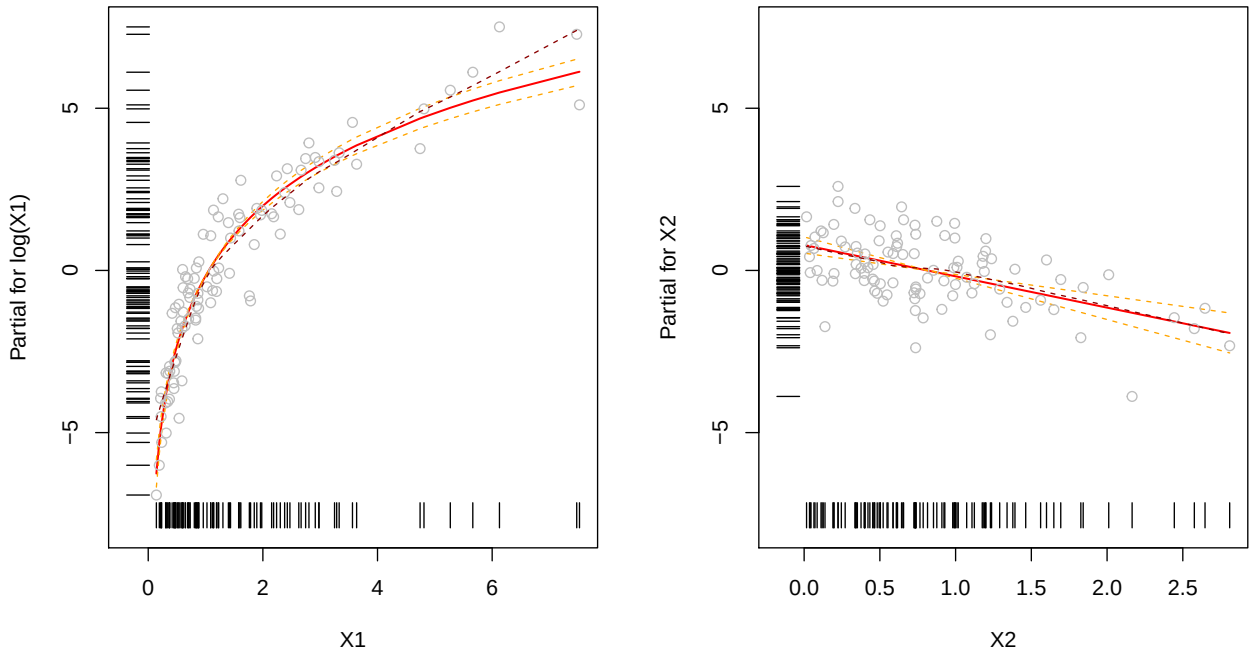
The lefthand side takes the response and removes the overall mean and the part of the response explained by X_2. The righthand side is the partial residual for X_1 (equal to the *term* for X_1 plus the residual e): it is the part of residual variation that is not explained by X_2 and that potentially can be explained by X_1. Equivalently, the partial residual for X_2 corresponds to

$$\widehat{\beta}_2(X_2 - \overline{X}_2) + e.$$

The argument `smooth=panel.smooth` in term plot draws a smooth lowess line through the points. This can help one assess departures from linearity.

Term plots with transformation of Y and X_1

```
par(mfrow = c(1,2))
termplot(lm(log(Y) ~ log(X1) + X2, data=simdat),
         partial.resid = T, se=T, rug=T, smooth = panel.smooth)
```



The model

$$\log(Y) = \widehat{\beta}_0 + \widehat{\beta}_1 \log(X_1) + \widehat{\beta}_2 X_2 + e$$

is equivalent to centered model

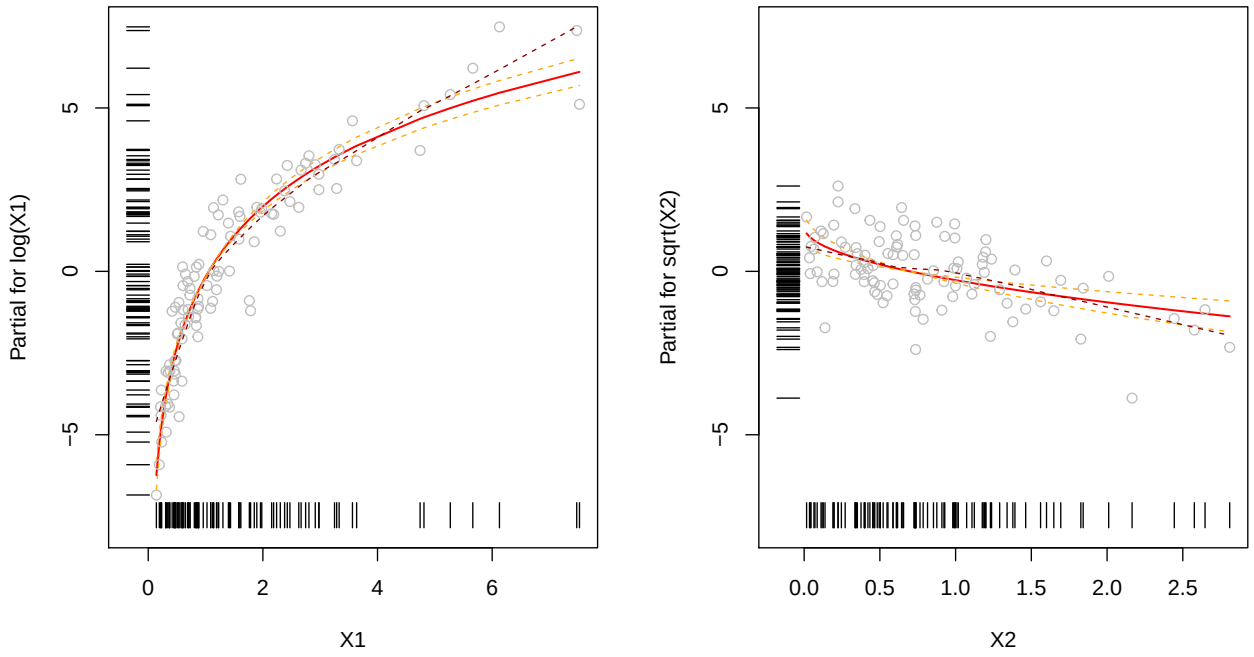
$$\log(Y) = \overline{\log(Y)} + \widehat{\beta}_1(\log(X_1) - \overline{\log(X_1)}) + \widehat{\beta}_2(X2 - \overline{X2}) + e$$

What are the terms and the partial residuals for X_1 and X_2 for the model with transformations of Y and X_1?

- $\widehat{\beta}_1(\log(X_1) - \overline{\log(X_1)})$ is the term for X_1, and $\widehat{\beta}_1(\log(X_1) - \overline{\log(X_1)}) + e$ is the corresponding partial residual
- $\widehat{\beta}_2(X_2 - \overline{X_2})$ is the term for X_2, and $\widehat{\beta}_2(X_2 - \overline{X_2}) + e$ is the corresponding partial residual

Term plots with transformation of Y, X_1, and X_2

```
par(mfrow = c(1,2))
termplot(lm(log(Y) ~ log(X1) + sqrt(X2), data=simdat),
         partial.resid = T, se=T, rug=T, smooth = panel.smooth)
```



Exercise

10. Examine added variable plots and term plots for your model above (UN data). Is it likely that any of the localities are influential for any of the terms? Which localities? Which terms?