Tutorial 9

YANG YANG

The Australian National University

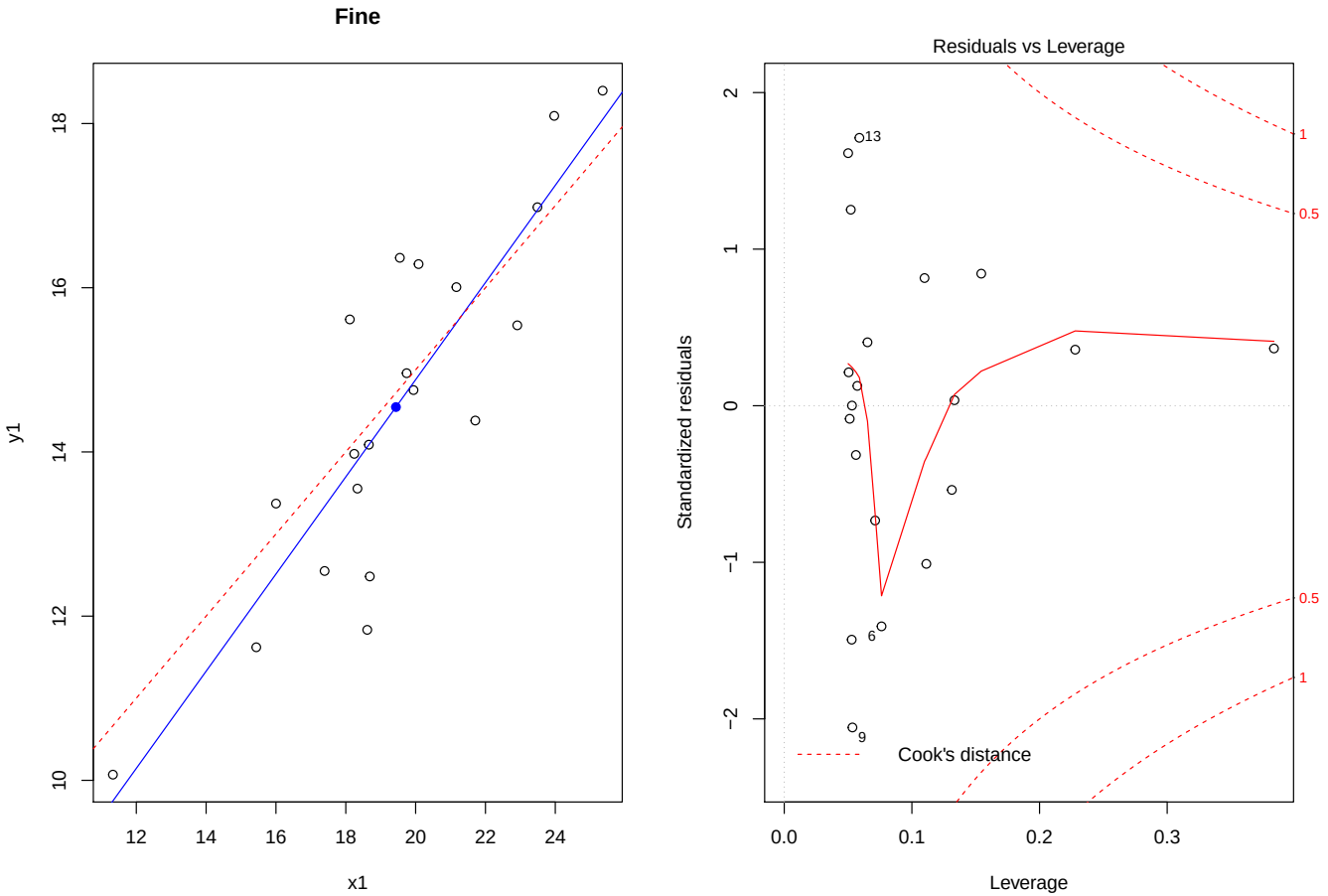
Week 10, 2017

Overview

Residual vs Leverage plot

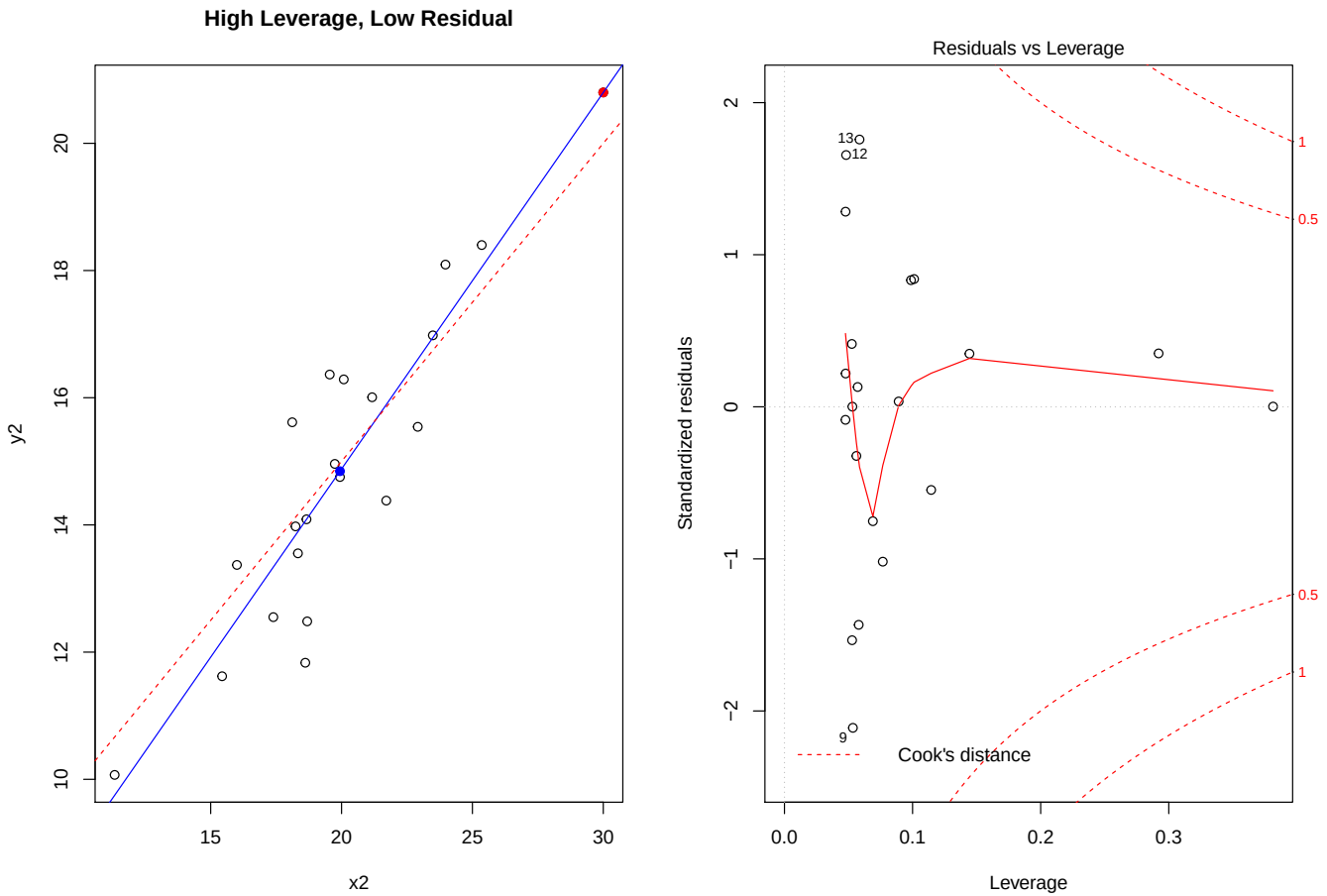
```
1 set.seed(20)
2
3 x1 = rnorm(20, mean=20, sd=3)
4 y1 = 5 + .5*x1 + rnorm(20)
5
6 x2 = c(x1, 30);          y2 = c(y1, 20.8)
7 x3 = c(x1, 19.44);       y3 = c(y1, 20.8)
8 x4 = c(x1, 30);          y4 = c(y1, 10)
9
```

Residual vs Leverage plot



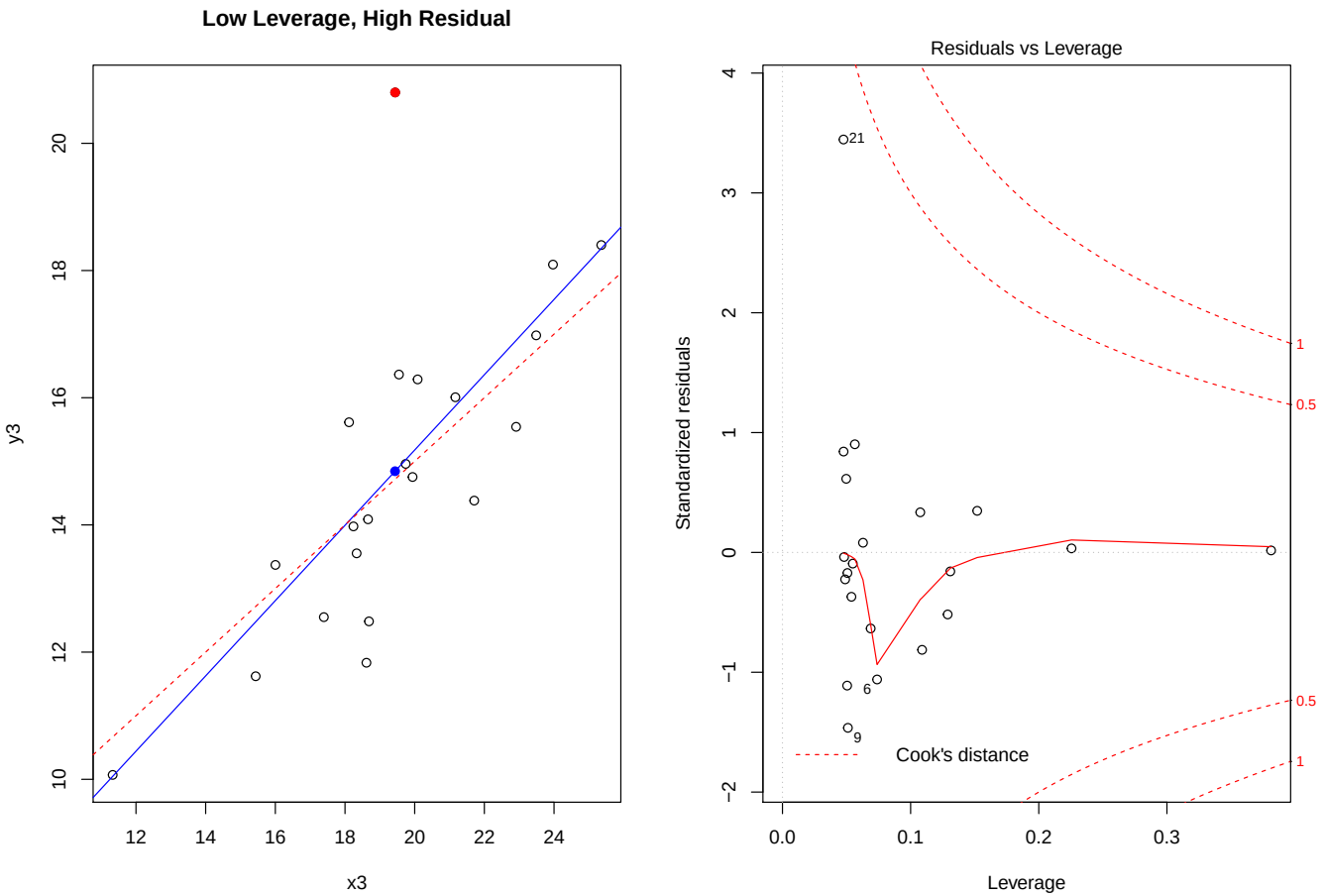
blue line = fitted model; red line = data generating process

Residual vs Leverage plot



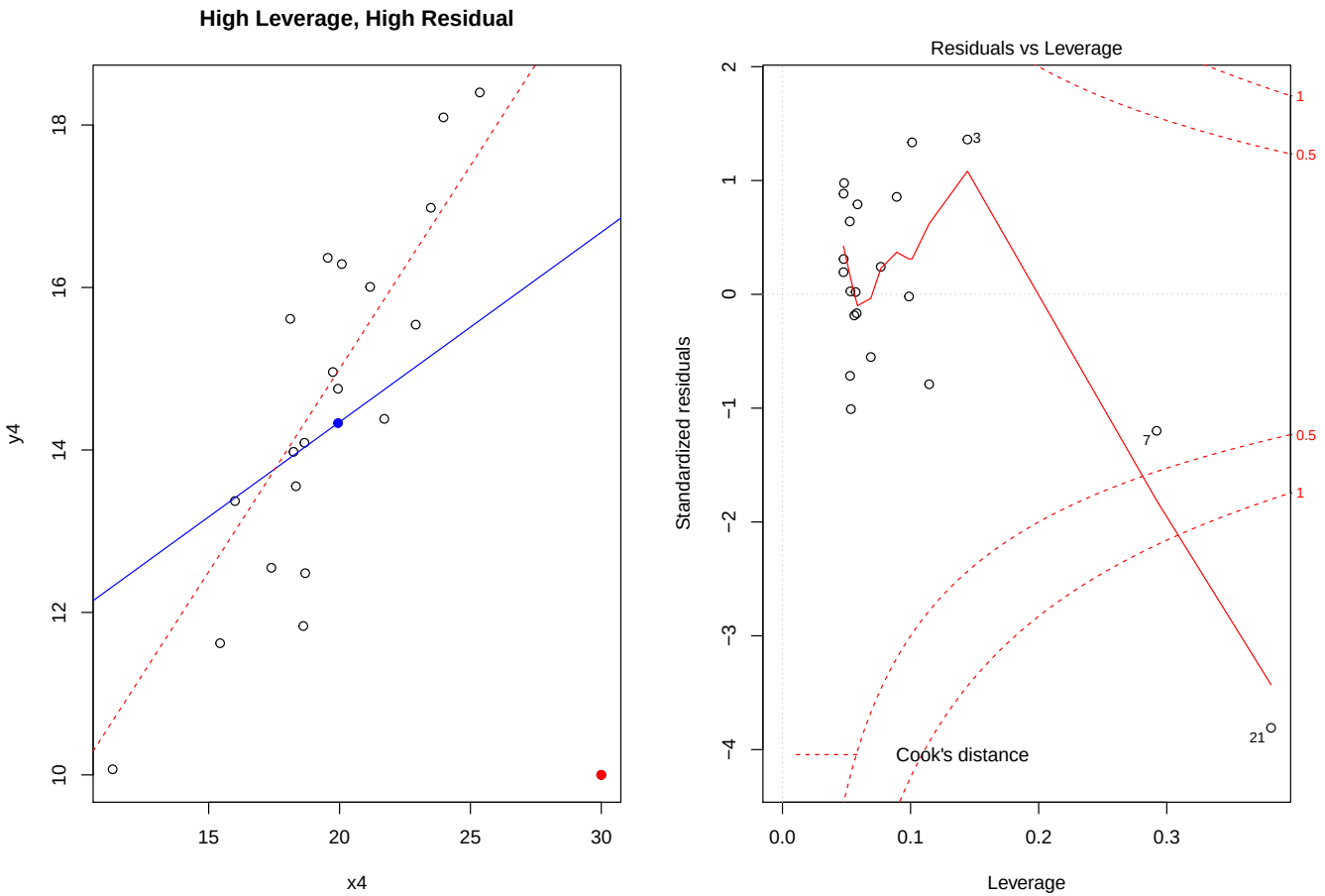
blue line = fitted model; red line = data generating process

Residual vs Leverage plot



blue line = fitted model; red line = data generating process

Residual vs Leverage plot



blue line = fitted model; red line = data generating process

How to select predictors? (Q4. (c))

If we are given a big dataset with many variables, we want to firstly narrow down the range of predictors.

1. Use `pairs(dataset)` and `cor(dataset)` to check the general structure of dataset. In particular,
   - **Search for predictor pairs which have strong correlations.** No more than one from each pair will be used in the model; otherwise, there is the problem of multicollinearity.
   - **Find predictors that have small correlation with the response.** We can later put such predictors last in the model to confirm that we can safely exclude them from our final model.

How to select predictors? (Q4. (d))

2. Build first model with all variables but put unlikely ones in the last. Check `anova(model)` and `summary(model)` to ensure we have p-values greater than 0.05 (or α) for those useless predictors.
    - Sequence of unlikely predictors generally will not matter in this step.
    - Refit the model without these unlikely predictors and compare the resulting coefficients, R^2 and MSE to those of the full regression.

How to select predictors? (Q4. (a))

3. Fit a multiple regression of the remaining variables. Examine the internally Studentized residuals versus fitted values as well as versus each of the predictors individually. Check the normal Q-Q plot.
   - The purpose is to check assumptions (i.e., constant variance, independence and normality) and find potential outliers.
   - If obvious problems found then we probably need to do transformations to some of predictors.

How to select predictors? (Q4. (b)&(F))

3. Build possible added variable plots and confirm linear structures in each plot.
   - If a non-linear relation is detected, try transformation to the predictor.
   - Write down potential influential outliers.
4. Do appropriate partial F-tests to confirm that all selected predictors are significant.
5. Calculate the leverages, DFFITS, DFBETAS and Cook's Distances for each of the data points. Check influence measures of the potential influential outliers noted in Step 3.
   - Remove influential outliers (if allowed!) and then check `anova(model)` and `summary(model)` to confirm improvement.

Q4 Hints

(a) **Internally** Studentized residuals versus fitted values plots; still use ± 2 to find vertical outliers

(b) Fit the predictor in question as the last explanatory variable before conducting partial F-tests.

(c) correlation matrix using `cor(savings[,3:6])`

(d) Remove a predictor, the R^2 must decrease but the adjusted-R^2 may increase.

(f) Interpret influence statistics in relative terms or with respect to cut-off values.

Added variable plot

Removes the effects of all the **other variables** from both the **response** and the particular **predictor in question** and then examine the relationship between the **remaining "unexplained" portions** of the two variables in question.
$\longrightarrow$ Remove any of the possible (linear) **confounding** effects of the **other variables**.

Added variable plot

How to **construct** an added variable plot?

- The residuals from regressing Y against all predictors other than $X_{interested}$ go on the vertical axis, while the residuals from regression $X_{interested}$ against all other predictors go on the horizontal axis.
- Since the mean residual from both of these regressions is zero, the mean point of ($X_{interested}$ given others, Y given others) will just be $(0, 0)$ which explains why the regression line in the added variable plot always goes through the origin.

Added variable plot

How to **interpret** an added variable plot?

- If an added variable plot shows a **linear structure**, this is evidence that the predictor variable under investigation should indeed be **included** in the model.
- On the other hand, if the added variable plot appears to be a **simple random scatter of points**, then we will likely conclude that the predictor is **not adding** any further explanation.
- We can include the calculated correlation coefficient `cor(Response,Predictor)` in the subtitle `sub = paste("r=",cor(Response, Predictor))` of the added variable plot.

Added variable plot

How to **interpret** an added variable plot?

- If the predictor x_i is truly related to the response variable, then the **added variable plot** should look like a **straight line** through the origin.
- **Curvature** in the added variable plot indicates a **non-linearly** in the relationship between the response and the predictor.
- Plots of the residuals vs each of the predictors are used to identify the probable source of any non-linearity.