1 Some Important Definitions

1.1 Homoscedasticity

- ***Homoscedascity*** is the technical term to describe the variance of the residuals being constant across the range of predicted values.
- ***Heteroscedascity*** is the converse scenario : the variance differs along the range of values.

Suppose you plot the individual residuals against the predicted value, the variance of the residuals predicted value should be constant.
Consider the red arrows in the picture below, intended to indicate the variance of the residuals at that part of the number line. For the OLS summption to be valid , the length of the red lines should be more or less the same.

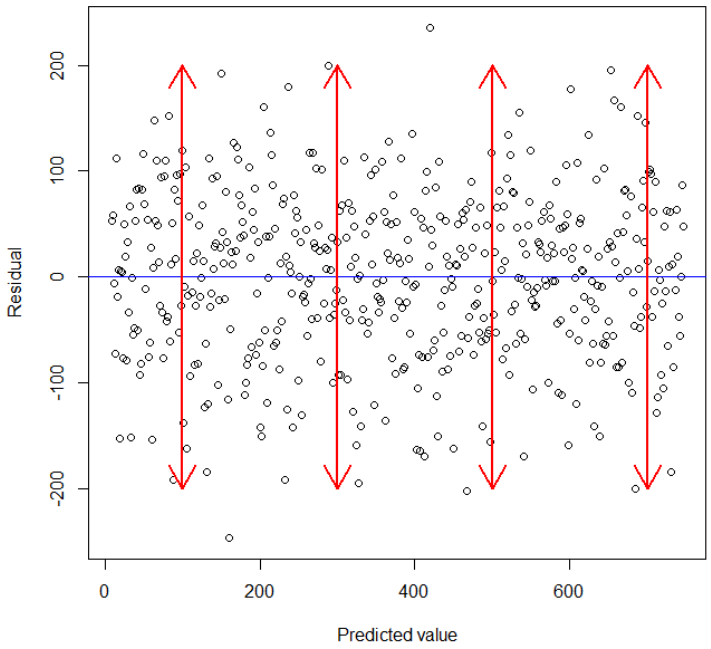

Figure 1:

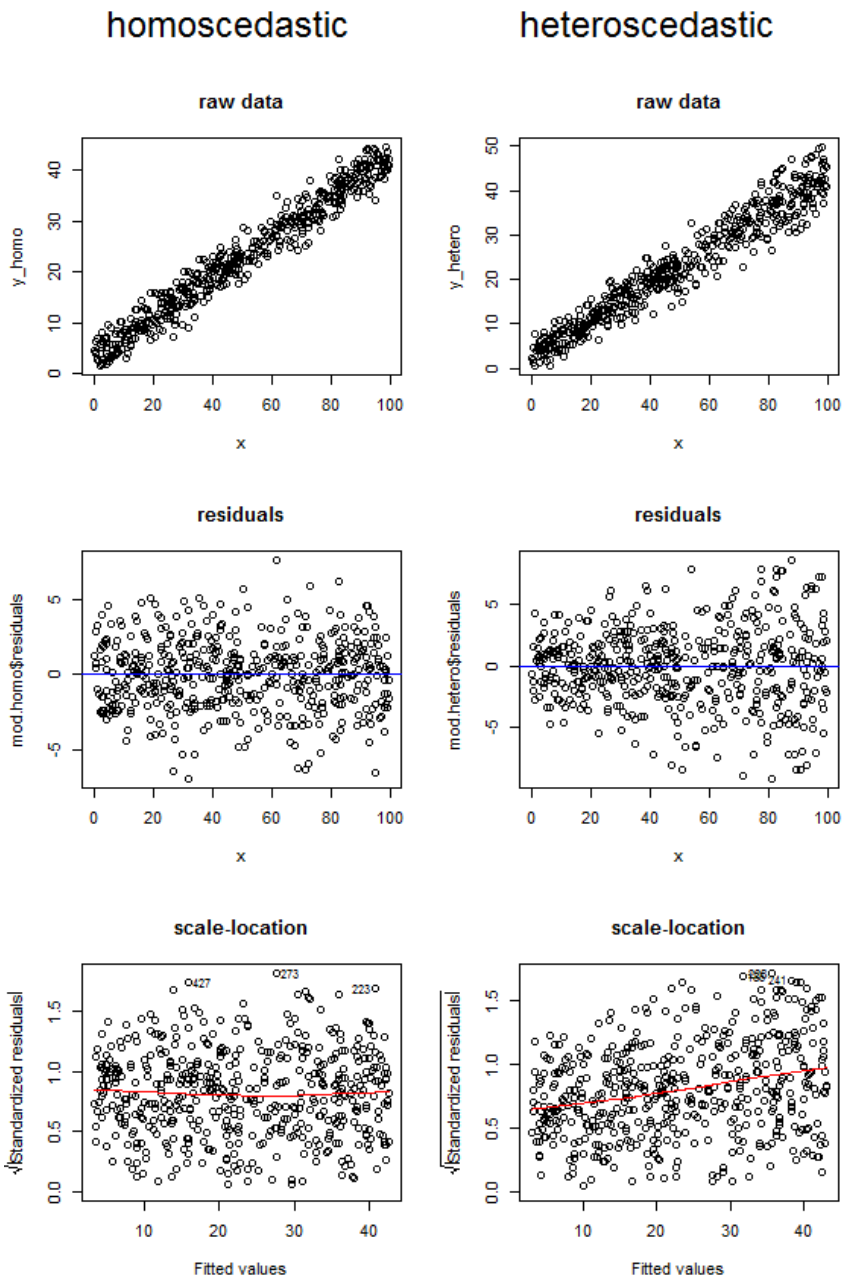
homoscedastic
heteroscedastic
raw data
raw data
y_homo
y_hetero
x
x
residuals
residuals
mod.homo$residuals
mod.hetero$residuals
x
x
scale-location
scale-location
√|Standardized residuals|
√|Standardized residuals|
427
273
223
Fitted values
Fitted values

Figure 2:

1.2 More Definitions

To understand a diagnostic plot called the residual-leverage plot, we need to understand three things:

- Leverage,
- Sstandardized residuals, and
- Cook's distance.

1.2.1 Leverage

To understand leverage, recognize that *Ordinary Least Squares* regression fits a line that will pass through the centre of your data, $(\bar{x}, \bar{y})$. The line can be shallowly or steeply sloped, but it will pivot around that point like a lever on a fulcrum.

This analogy can be taken fairly literally: because OLS seeks to minimize the vertical distances between the data and the line, the data points that are further out towards the extremes of X will push / pull harder on the lever (i.e., the regression line); they have more leverage. One result of this could be that the results you get are driven by a few data points; that these diagnostic plots are intended to identify.

1.2.2 Standardization

Another result of the fact that points further out on X have more leverage is that they tend to be closer to the regression line (or more accurately: the regression line is fit so as to be closer to them) than points that are near $\bar{x}$. In other words, the residual standard deviation can differ at different points on X (even if the error standard deviation is constant). To correct for this, residuals are often standardized so that they have constant variance (assuming the underlying data generating process is homoscedastic, of course).

1.2.3 Cook's Distance

One way to think about whether or not the results you have were driven by a given data point is to calculate how far the predicted values for your data would move if your model were fit without the data point in question.

This calculated total distance is called **Cook's distance**. Fortunately, you don't have to rerun your regression model N times to find out how far the predicted values will move, Cook's D is a function of the leverage and standardized residual associated with each data point.

With these facts in mind, consider the plots associated with four different situations:

1. a dataset where everything is fine
2. a dataset with a high-leverage, but low-standardized residual point
3. a dataset with a low-leverage, but high-standardized residual point
4. a dataset with a high-leverage, high-standardized residual point

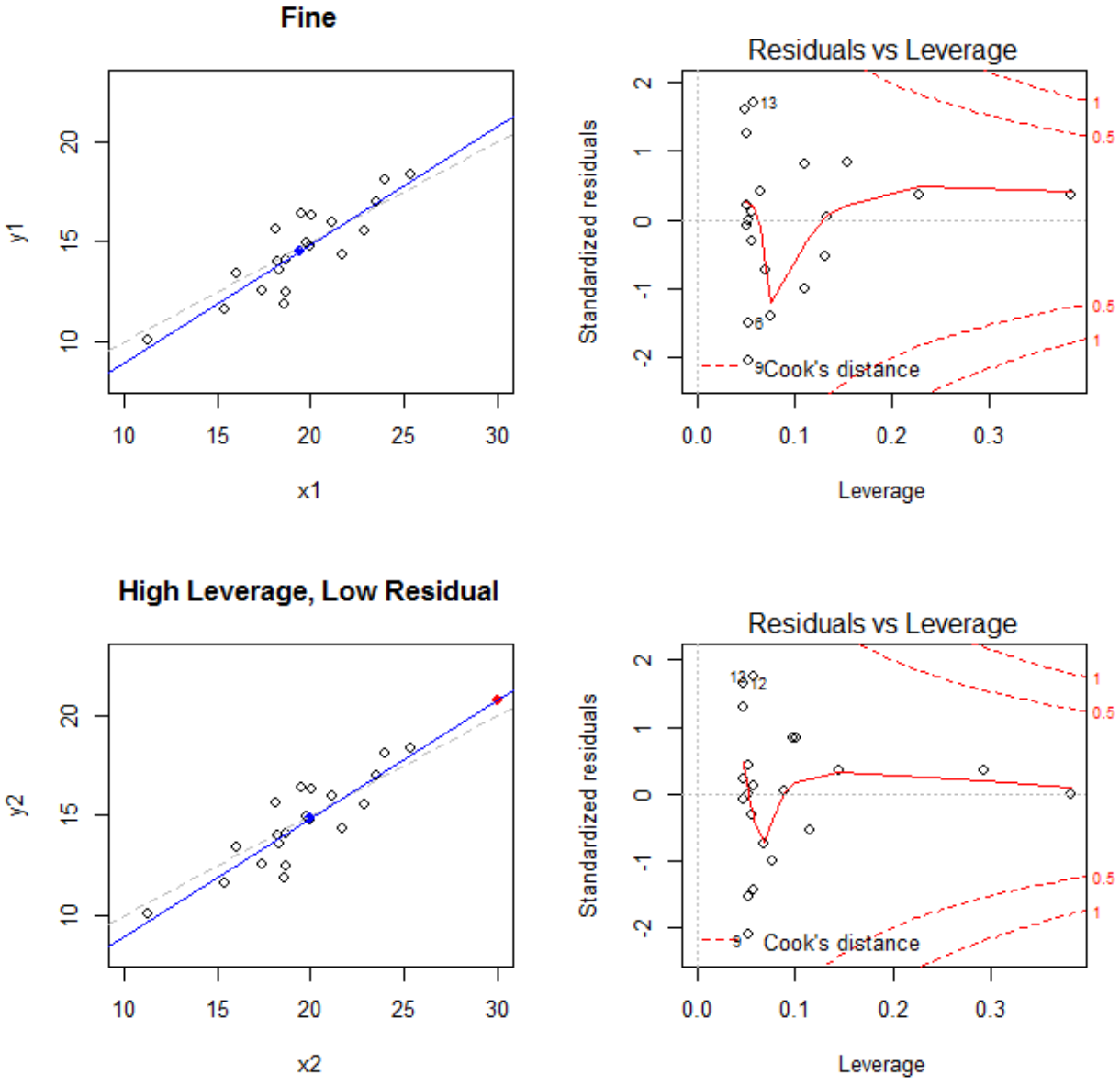
Fine
y1
x1
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
High Leverage, Low Residual
y2
x2
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance

Figure 3:

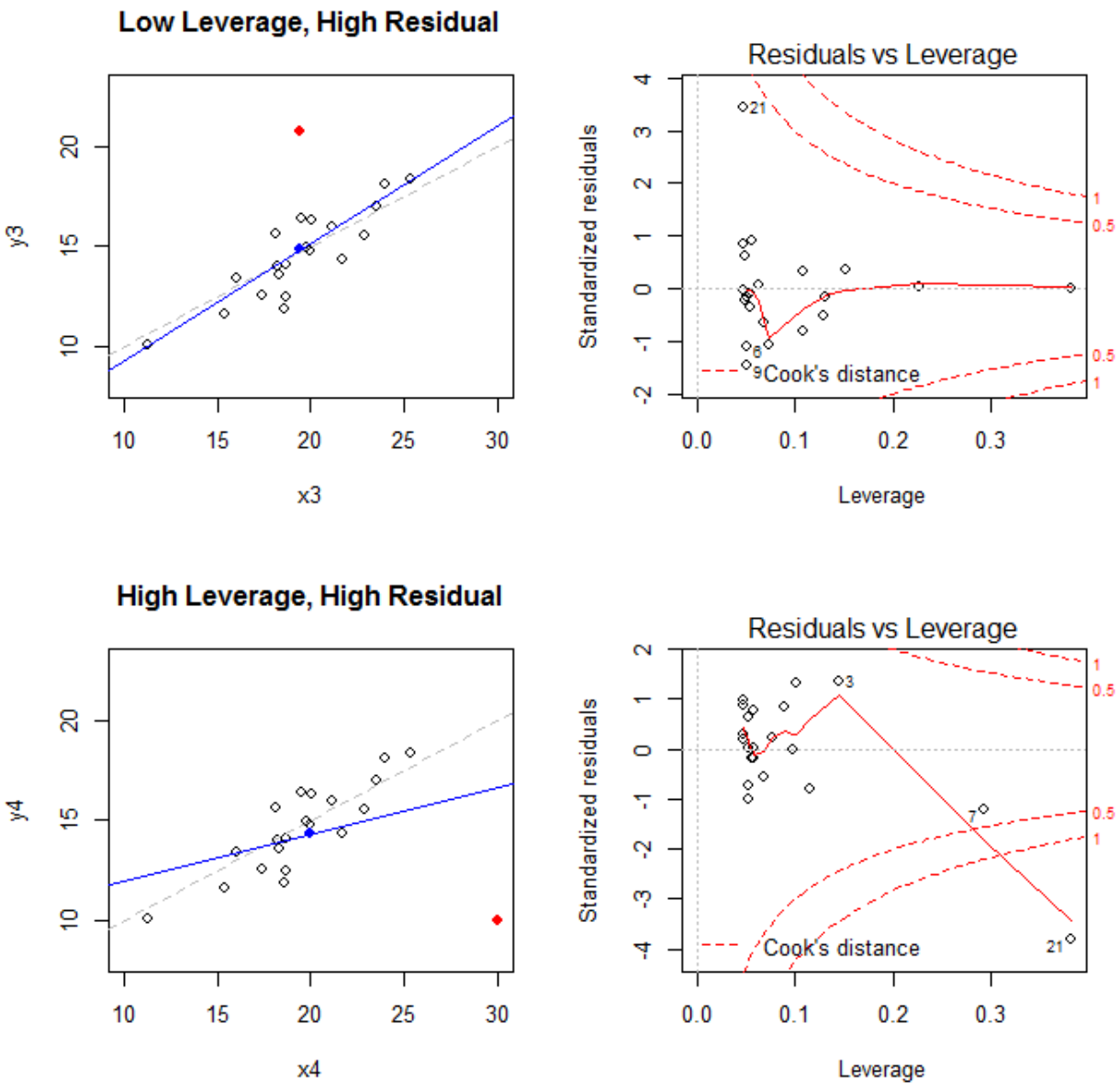
Low Leverage, High Residual
y3
x3
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
High Leverage, High Residual
y4
x4
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance

Figure 4:

- The plots on the left show the data, the center of the data with a blue dot, the underlying data generating process with a dashed gray line, the model fit with a blue line, and the special point with a red dot.
- On the right are the corresponding residual-leverage plots; the special point is 21.
- The model is badly distorted primarily in the fourth case where there is a point with high leverage and a large (negative) standardized residual.