Brief Introduction to R

1. simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + e_i \text{ where } e_i \text{ are i.i.d from } N(0, \sigma_e^2).$$

   (a) model assumptions
   (1) The data follow a straight line
   (2) e_i are independent
   (3)Errors have constant variance
   (4) Errors follow a normal distribution

   (b) $\hat{\beta}_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$
   $s_e^2 = MSError = \sum(y_i - \hat{y}_i)^2/(n-2)$
   $se(\hat{\beta}_1) = \frac{s_e}{\sqrt{\sum(x_i-\bar{x})^2}}$, $se(\hat{\beta}_0) = s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}}$
   $se(\hat{Y}_{est}) = s_e\sqrt{\frac{1}{n} + \frac{(x_*-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$, $se(\hat{Y}_{pred)=}s_e\sqrt{1 + \frac{1}{n} + \frac{(x_*-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$

2. Assessing Assumptions

   (a) Types of residuals
   (1) raw residuals $r_i = y_i - \hat{y}_i$, where $\sum r_i = 0$, and $SSError = \sum r_i^2$
   (2) internally studentized residuals $rint_i = \frac{y_i-\hat{y}_i}{s_e\sqrt{1-h_i}}$, where $h_i = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}$
   (3) externally studentized residuals $rext_i = \frac{y_i-\hat{y}_i}{s_{e(i)}\sqrt{1-h_i}}$

   (b) Outlier Test
   (1) Delete the questionable observation(denoted by x_*) and refit the model using the reduced data set.
   (2) Obtain $\hat{Y}_{pred}$ and $se(\hat{Y}_{pred})$
   (3) Do the test: $H_0 : Y_{observed} = Y_{predicted}$ vs. $H_A : Y_{observed} \neq Y_{predicted}$

$$T = \frac{Y_{obs} - \hat{Y}_{pred}}{se(\hat{Y}_{pred})}$$

   (4) Compute two-sided p-value based on T-distribution with df=(n-3).

   (c) Influential Observations
   (1) leverage $h_i = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}$, if $h_i > 4/n$, we consider it as influential.
   (2) Cook's distance $D_i = \frac{\sum_{j=1}^{n}(\hat{y_j} - \hat{y_{j(i)}})^2}{2s_e^2}$, if $D_i > 1$, we consider it as influential.

Practice Problems

1. Consider the following (artificial) data set.

x:	1	5	6	7	8	9	10
y:	1	11.7	11.8	9.7	8.5	8.6	7.3

(1) Fit the data using simple linear regression.
(2) Is it a good fit? Remove any possible outlier you think and then refit the data.
(3) What can you conclude by comparing the two fits?
(4) Based on (1), which observations can be considered as influential points?

2. Consider the following (artificial) data set.

x:	1	2	3	4	5	6	7
y:	10.9	6.1	2.8	1.9	3.1	6.0	11.2

(a) Plot y vs. x.

(b) Fit the data using simple linear regression. How good is your fit?

(c) Plot the residuals vs. y

(d) Plot the residuals vs. x.

(e) Plot the residuals vs. $\hat{y}$.

(f) Compare the plots from (c), (d) and (e). What do you conclude? Which plot provides a better indication of the lack of fit?