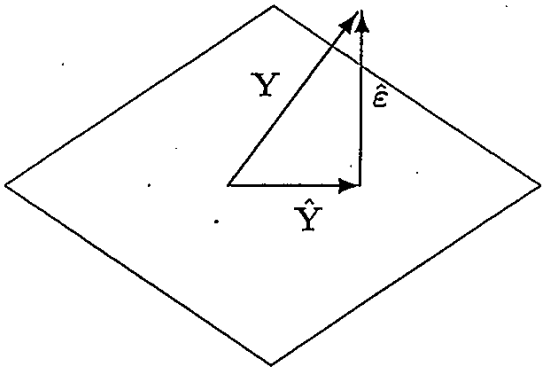
Lab 3: Linear Regression

1 Matrix representation for regression model

$$\begin{aligned} Var(\mathbf{Y}) &= E\big[\mathbf{Y} - E(\mathbf{Y})\big]\big[\mathbf{Y} - E(\mathbf{Y})\big]^T = \Sigma \\ Var(\mathbf{AY}) &= \mathbf{A}\Sigma\mathbf{A}^T \end{aligned} \tag{1}$$

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \epsilon \\ \epsilon &\sim (\mathbf{0}, \sigma^2\mathbf{I}) \end{aligned} \tag{2}$$

The least square estimate $\beta_{lse} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is unbiased. $Var(\beta_{lse}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$. It also has nice geometric interpretation.



With the assumption that ϵ is normally distributed, we have

$$\beta_{lse} \sim N_p(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

2 Leverage

In simple linear regression $Y_i = \alpha + \beta X_i + e_i, \quad E(\epsilon_i) = 0$, we have

$$\begin{aligned} \hat{\beta} &= \frac{\sum(Y_i - \bar{Y})(X_i - \bar{X})}{\sum(X_i - \bar{(X)})^2} = \hat{\rho}_{XY}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \\ &= \frac{\sum \frac{(Y_i - \bar{Y})}{(X_i - \bar{X})}(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2} = \sum w_i\hat{\beta}_i \\ w_i &= \frac{(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2}, \text{ leverage} \\ \hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X} \end{aligned} \tag{3}$$

Generally, we have

$$\begin{aligned} \widehat{\mathbf{Y}} &= \mathbf{X}\hat{\beta} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \mathbf{HY} \end{aligned} \tag{4}$$

H is called the hat matrix. And

$$\begin{aligned} E(\hat{\epsilon}) &= 0 \\ Var(\hat{\epsilon}) &= \sigma^2(\mathbf{I}-\mathbf{H}) \\ i.e.,\ Var(\hat{\epsilon}_i) &= \sigma^2(1-h_{ii}) \\ h_{ii} &= x_i^{;}(\mathbf{X}^T\mathbf{X})^{-1}x_i. \end{aligned} \tag{5}$$

3 Regression diagnose

- **R-Squared**

$$R_2 = \frac{SSR}{SST} \tag{6}$$

- **Scatterplots**
  pair-wise relation among variables

- **Residual plots(rvfplot)**
  Examine the independence and normality assumption of the error term. Also check the appropriateness of the model.

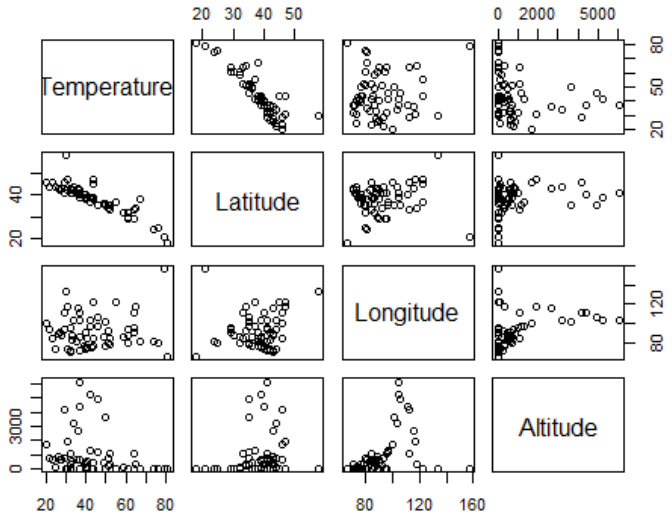
- **Adjusted variable plots (Avplots)**
  Suppose the model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$. We want to demonstrate the relation between dependent variable Y and a single covariate X_1, while adjusted for other covariates. It can also help to find the influential point in the particular variable.

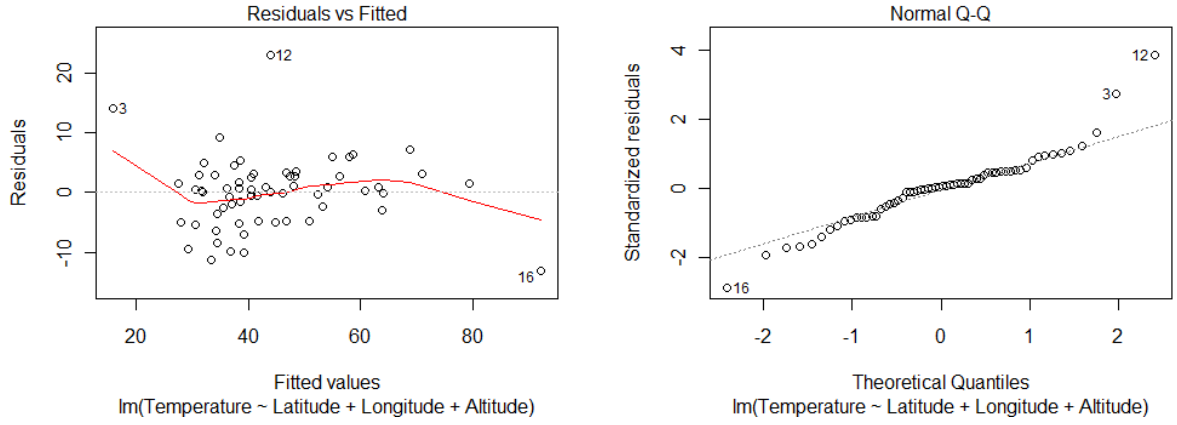
  (a). Regress Y on X_2 and get residual $e(Y|X_2)$;

  (b). Regress X_1 on X_2 and get the residual $e(X_1|X_2)$

  (c). Plot $e(Y|X_2)$ against $e(X_1|X_2)$

  It turns out that the slope of $e(Y|X_2)$ $e(X_1|X_2)$ is the same as the slope for X_1 in the original MLR model.

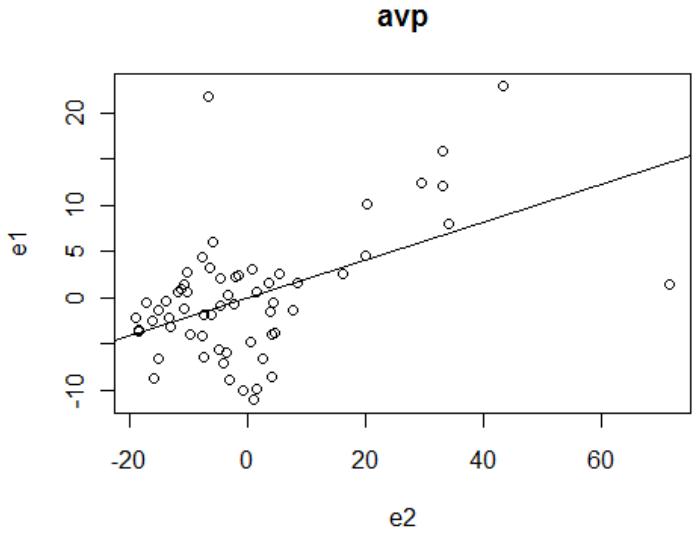




4 ANCOVA

How much the association of Y and X_1 differs across levels of X_2?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$


avp
e1
e2
-10
0
5
10
20
-20
0
20
40
60