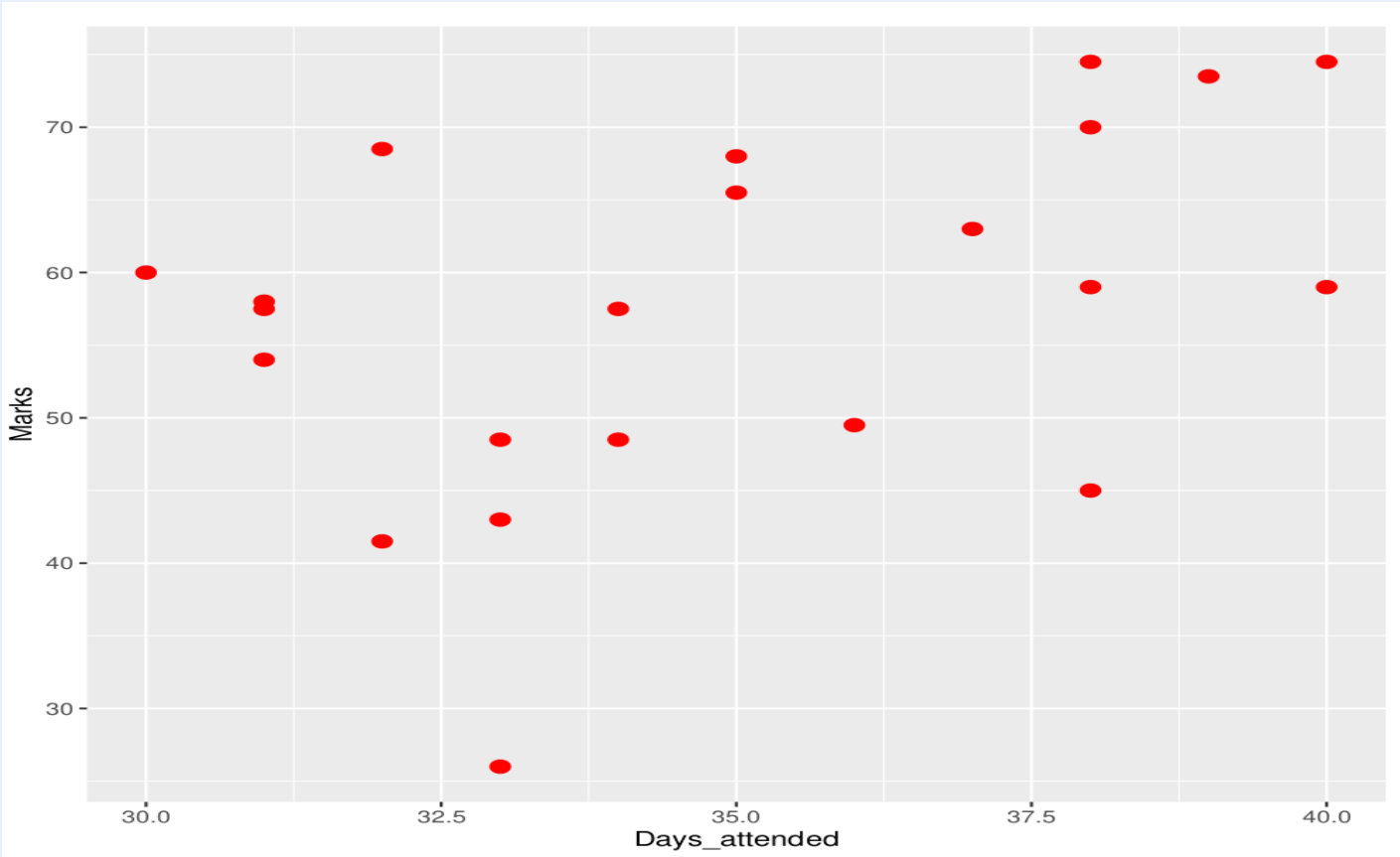
Statistical Inference (MA862)

Lecture Slides

Topic 5: Linear Regression

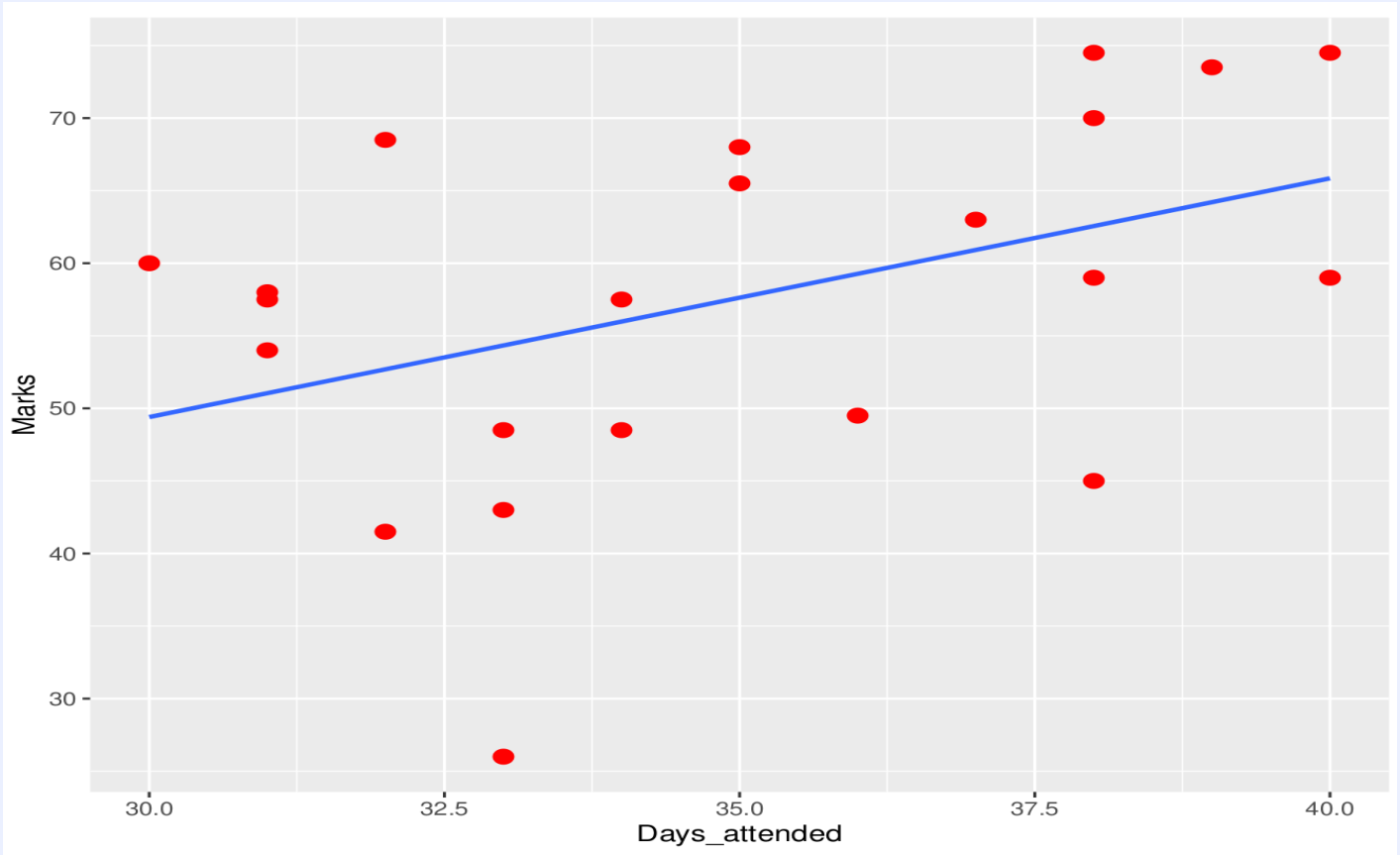
Regression

- Question: What is the impact of attending classes on students' final marks?
- Let's start with a real data from IITG which you can feel about it!!



Regression

- Question: What is the impact of attending classes on students' final marks?
- Let's start with a real data from IITG which you can feel about it!!



Linear Regressions

- We have one particular variable that we are interested in understanding or modeling, such as sales of a particular product, sale price of a home, or voting preference of a particular voter. This variable is called the target, response, or dependent variable, and is usually represented by y.
- We have a set of p other variables that we think might be useful in predicting or modeling the target variable (for *e.g.* the price of the product, the competitor's price, and so on; or the lot size, number of bedrooms, number of bathrooms of the home, and so on; or the gender, age, income, party membership of the voter, and so on). These are called the predicting, independent variables, or features and are usually represented by $x_1, x_2, \ldots, x_p$.

Linear Regressions

- Thus, we have

$$y = f(\boldsymbol{x}; \boldsymbol{\beta}) + \varepsilon,$$

  for some real valued function f, where $\boldsymbol{x}$ is vector of predictors, $\boldsymbol{\beta}$ is the vector of parameters, and ε is error.
- If f is linear in the parameters vector $\boldsymbol{\beta}$, then the regression is called linear regression.

Examples and Role of Transformations

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ is a linear model.
- $y = \beta_0 + \beta_1 x + \beta_2 x^2$ is a linear model, because it is linear in β (even though not in x).
- $y = \beta_0 + \beta_1 x^{\beta_2}$ is a non-linear model, as it is not linear in β.
- $y = \beta_0 x^{\beta_1}$ is not a linear model, but $\ln y = \ln \beta_0 + \beta_1 \ln x$ is.
- $y = \frac{e^{\beta x}}{1+e^{\beta x}}$, where $y \in (0, 1)$
- $y = \frac{1}{\beta_0+\beta_1 x_1+\beta_2 x_2}$.

Main use of Linear Regressions

Typically, a regression analysis is used for one (or more) of three purposes:

1. modeling the relationship between x and y;
2. prediction of the target variable (forecasting);
3. testing of hypotheses.

Simple Linear Regression

- Just one predictor x, i.e. $p = 1$.
- The model for the simple linear regression is given by

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where y is the outcome variable (random), x is the independent/predictor variable (non-random) and ϵ is the random error term. β_0 (intercept) and β_1 (slope) are model parameters (unknown constants).

- Equivalently, the model can be written for $i = 1, 2, \cdots, n$ number of observations $(x_1, y_1), \cdots, (x_n, y_n)$ as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n.$$

- How do you interpret β_0 (intercept) and β_1 (slope)?

Least Squares Estimation

- Goal: To estimate β_0, β_1 by minimizing error in some sense (*e.g.* squared error)
- One reasonable way is to use the principle of Least Squares, *i.e.* minimize the objective function

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

  with respect to β_0, β_1.
- Differentiate $Q(\beta_0, \beta_1)$ with respect to β_0, β_1 and equate the partial derivatives to zero to get the estimates $\hat{\beta}_0, \hat{\beta}_1$.
- The resulting equations are called normal equations:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \text{ and } \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

Least Squares Estimation

- The solution is given by

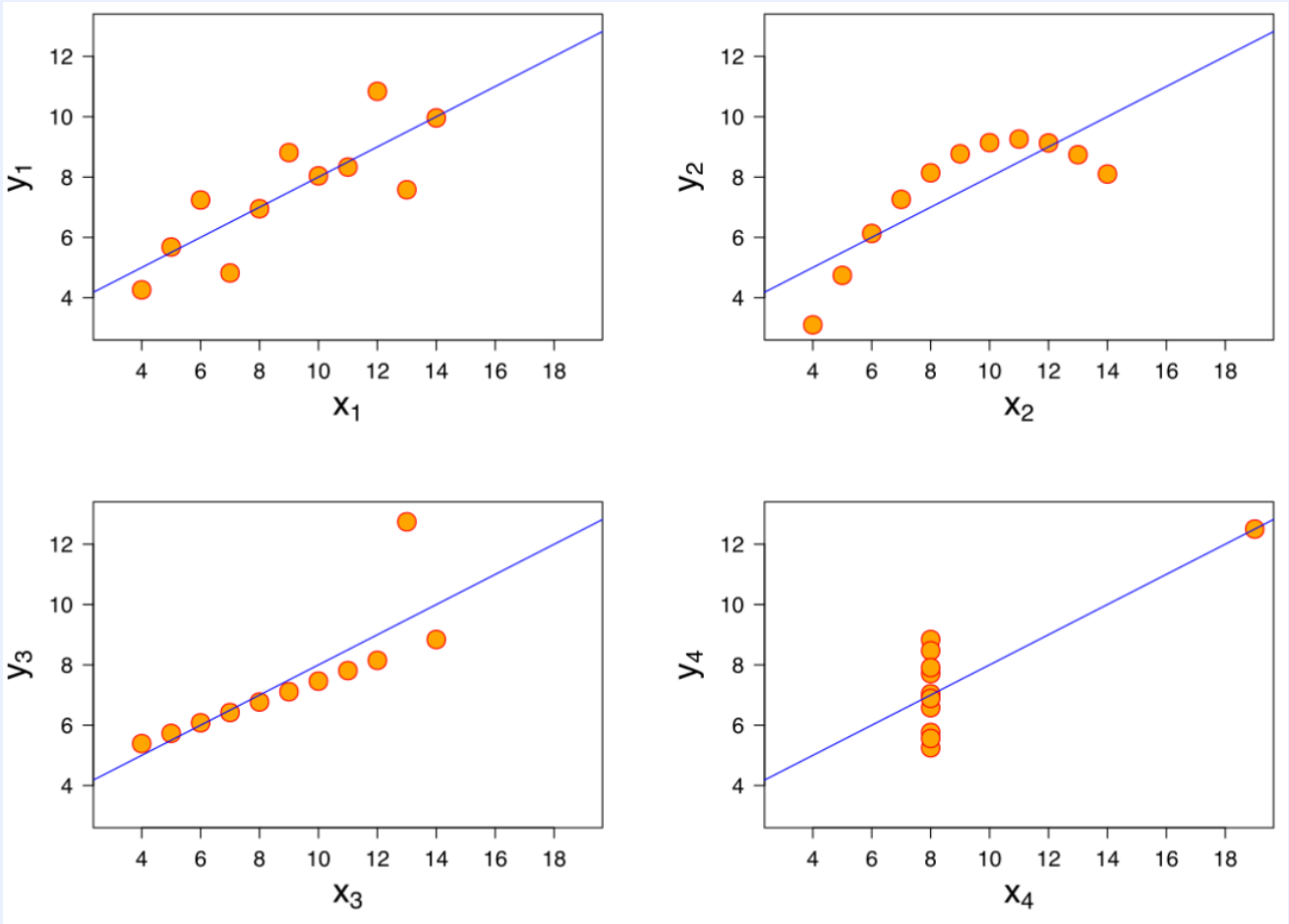
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = S_{xy}/S_{xx}$$

  where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{and} \quad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the least squares estimator (LSE) of β_0 and β_1, respectively.

Importance of graphing data before analyzing it



Which one of the above do you think has highest value of absolute correlation and ideal for linear regression?

Importance of graphing data before analyzing it

- In all the four graphs: mean of x $= 9$ (with variance 11); mean of y $= 7.50$ (with variance 4.1) ; correlation between x and y $=$ 0.816
- Fitted linear regression in each cases: $y = 3 + 0.5x$
- In 1973, Anscombe demonstrated the importance of graphing data before analyzing it and the effect of outliers on statistical properties

Assumptions

1. The errors are uncorrelated with each other, *i.e.*,

$$Cov(\epsilon_i, \epsilon_j) = 0 \text{ for } i \neq j.$$

2. The expected value of the errors is zero, *i.e.*,

$$E(\epsilon_i) = 0 \text{ for all } i.$$

3. The errors are homoscedastic (constant variance), *i.e.*,

$$Var(y_i) = \sigma^2 \text{ for all } i$$

.

Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combination of y_i's.
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased for β_0 and β_1, respectively.
- Variances of $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\begin{aligned} Var(\hat{\beta}_0) &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \\ Var(\hat{\beta}_1) &= \frac{\sigma^2}{S_{xx}} \end{aligned}$$

Gauss-Markov Theorem

Definition 5.1: An estimator $\hat{\theta}$ of θ is called linear estimator of θ if $\hat{\theta}$ is a linear combination of random observations.

Definition 5.2: An estimator $\hat{\theta}$ of θ is called the best linear unbiased estimator (BLUE) of θ if $\hat{\theta}$ is linear and unbiased estimator of θ and $\hat{\theta}$ has minimum variance among all linear unbiased estimator of θ.

Theorem 5.1: Under the above assumptions of linear regression, the LSEs $\hat{\beta}_0$ and $\hat{\beta}_1$ are BLUE of β_0 and β_1, respectively.

A Few Definitions

- Fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,\ i = 1, \ldots, n.$
- Residuals: $e_i = y_i - \hat{y}_i,\ i = 1, \ldots, n.$
- The objective function evaluated at the LSEs is called the residual sum of squares (SS_{Res}).

$$SS_{Res} = Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 .$$

- The following quantity is called total sum of squares (SS_T):

$$SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2 .$$

- The following quantity is called regression sum of squares (SS_{Reg}):

$$SS_{Reg} = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 .$$

Properties of Least Squares Fit

- $\sum_{i=1}^{n} e_i = 0.$
- $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i.$
- $\sum_{i=1}^{n} x_i e_i = 0.$
- $\sum_{i=1}^{n} \hat{y}_i e_i = 0.$
- $SS_T = SS_{Reg} + SS_{Res}.$

Estimation of Error Variance

- It can be shown that

$$E(SS_{Res}) = (n-2)\sigma^2.$$

- Hence, $\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res}$ is an unbiased estimator of σ^2. Here MS_{Res} is the residual mean square.
- Observed value of $\hat{\sigma}^2 = \frac{SS_{Res}}{n-2}$ is called Residual variance. It's square root is called the residual standard error.
- A convenient computing formula for SS_{Res} is

$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}.$$

Another Assumption

- Errors (ϵ_i) are normally distributed

This assumption is needed for further analysis – Hypothesis testing, construction of confidence intervals.

Hypothesis Testing: β_1

- Want to test the hypothesis that the slope parameter (β_1) equals to a constant (a value, say β_{10}):

$$H_0 : \beta_1 = \beta_{10} \text{ ag. } H_1 : \beta_1 \neq \beta_{10}$$

- Note that, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and y_i's are independent.
- $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}}) \implies z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$. But σ is unknown.
- $\frac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$. Also MS_{Res} and $\hat{\beta}_1$ are independent.
- Therefore, the test statistic is

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} \sim t_{n-2}, \text{ under } H_0.$$

- Reject H_0 iff $|t| > t_{n-2,\alpha/2}$; (at level α).

F-Test for Regression

- To test

$$H_0 : E(Y|x) = \beta_0 \text{ ag. } H_1 : E(Y|X = x) = \beta_0 + \beta_1 x$$

- Test statistics is a ratio, defined as F :

$$F = \frac{SS_{Reg}/1}{\hat{\sigma}^2} = \frac{SS_{Reg}/1}{SS_{Res}/(n-2)} \sim F_{1,n-2},$$

where $SS_{Reg} = \sum_i (\hat{y}_i - \bar{y})^2$

Hypothesis Testing: β_0

- Want to test the hypothesis that the **intercept** parameter (β_0) equals to a constant (a value, say β_{00}):

$$H_0 : \beta_0 = \beta_{00} \text{ ag. } H_1 : \beta_0 \neq \beta_{00}$$

- $\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})) \implies z = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}} \sim N(0,1)$. But σ is unknown.
- $\frac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$. Also MS_{Res} and $\hat{\beta}_1$ are independent.
- Therefore, the test statistic is

$$t = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}} \sim t_{n-2}, \text{ under } H_0.$$

- Reject H_0 iff $|t| > t_{n-2,\alpha/2}$; (at level α).

Interval Estimation: β_0 and β_1

- To get the CI for β_0 and β_1, the pivots are

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{MS_{Res}(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}}, \text{ and } \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_{Res}/S_{xx}}}, \text{ respectively.}$$

- A $100(1-\alpha)\%$ CI for β_0 is

$$\left[\hat{\beta}_0 \mp t_{n-2,\alpha/2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}\right].$$

- A $100(1-\alpha)\%$ CI for β_1 is

$$\left[\hat{\beta}_1 \mp t_{n-2,\alpha/2}\sqrt{\frac{MS_{Res}}{S_{xx}}}\right].$$

Interval Estimation: CI for σ^2

- To get the CI for σ^2, the pivots is

$$\frac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$$

- A $100(1-\alpha)\%$ CI for σ^2 is

$$\left[\frac{(n-2)MS_{Res}}{\chi^2_{n-2;\alpha/2}}, \frac{(n-2)MS_{Res}}{\chi^2_{n-2;1-\alpha/2}}\right].$$

Interval Estimation: CI for mean response

- A regression model can be used to estimate the mean response $E(y)$ for a particular value of the regressor variable x. Let x_0 be a value of the regressor variable. Then $E(y|x_0) = \beta_0 + \beta_1 x_0$.
- Then, $\widehat{y_0} = \widehat{E(y|x_0)} = \hat{\beta}_0 + \hat{\beta}_1 x_0$
- And, $\widehat{y_0} \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$
- Pivot: $\dfrac{\widehat{y_0} - (\beta_0 + \beta_1 x_0)}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$
- A $100(1-\alpha)\%$ CI for $\beta_0 + \beta_1 x_0$ is

$$\left[\widehat{y_0} \mp t_{n-2;\alpha/2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right].$$

Prediction Interval for New Observation:

- Let x_0 be a value of the regressor variable.
- The true value of the response is y_0 (corresponding to x_0).
- We want to provide an interval I such that $P(y_0 \in I) = 1 - \alpha$
- Note that the point estimate of y_0 is $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.
- Consider $\psi = y_0 - \hat{y}_0$.
- Then, $E(\psi) = 0$, $Var(\psi) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$
- $\psi \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$
- Pivot: $\frac{y_0 - \hat{y}_0}{\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$
- A $100(1 - \alpha)\%$ prediction interval is

$$\left[\hat{y}_0 \mp t_{n-2;\alpha/2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right].$$

Coefficient of determination: R^2

- Coefficient of determination is given by

$$R^2 = \frac{SS_{Reg}}{SS_T} = 1 - \frac{SS_{Res}}{SS_T} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

- It is a bounded quantity: $0 \leq R^2 \leq 1$.
- R^2 is interpreted as the proportion of variation explained by the model.
- Higher values of R^2 are desirable (R^2 close to 1 indicates a good fit).
- But "how high is high?": depends on the context.

Simple Linear Regression in Heights Data

- Data on heights of $n = 1375$ mothers in the UK under the age of 65 and one of their adult daughters over the age of 18 (collected and organized during the period 1893–1898 by the famous statistician Karl Pearson).
- A historical use of regression to study inheritance of height from generation to generation.
- Let's fit linear regression with this data set using `R`.

Multiple Linear Regression

- Data (of sample size n) looks like a matrix:

$$\begin{pmatrix} y_1 & x_{11} & x_{12} & \cdots & x_{1p} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ y_n & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Multiple Linear Regression

- In general, the response (y) may be related to p regressors (input variables/predictors).
- The model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, i = 1, \ldots, n$$

  is called multiple linear regression. A regression model with p regressors.
- The parameters $\beta_j, j = 0, 1, 2, \cdots, p$ are called regression coefficients.
- $\beta_j, j = 0, 1, 2, \cdots, p$ represents the change in the average value of the response for a unit change in j^{th} regressor keeping other regressors fixed.
- As before, ϵ_i's are i.i.d. with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$ for $i = 1, \ldots, n$

Multiple Linear Regression

- Then we can write the model in a more compact form:

$$\underline{y}_{n\times 1} = X_{n\times(p+1)}\underline{\beta}_{(p+1)\times 1} + \underline{\epsilon}_{n\times 1}$$

where

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \underline{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \underline{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix},$$

- X is called the design matrix

Multiple Linear Regression

- Multiple Linear Regression Model:

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}.$$

- $\underline{\epsilon}$ is a random vector.
- Assumptions: $E(\underline{\epsilon}) = 0$ and $Var(\underline{\epsilon}) = \sigma^2 I$ and all the assumptions stated in simple linear regression.

Estimation of Model Parameters

- The LSEs of $\beta_0, \beta_1, \cdots, \beta_p$ are

$$
\begin{aligned}
& \underset{\beta_0,\beta_1,\cdots,\beta_p}{\operatorname{argmin}} \; Q(\beta_0, \beta_1, \cdots, \beta_p) \\
= & \underset{\beta_0,\beta_1,\cdots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip} \right)^2 \\
= & \underset{\underline{\beta}}{\operatorname{argmin}} \; (\underline{y} - X\underline{\beta})^T (\underline{y} - X\underline{\beta}) \\
= & \underset{\underline{\beta}}{\operatorname{argmin}} (\underline{y}^T \underline{y} - 2\underline{\beta}^T X^T \underline{y} + \underline{\beta}^T X^T X \underline{\beta})
\end{aligned}
$$

Estimation of Model Parameters

- Differentiating $Q(\underline{\beta})$ with respect to $\underline{\beta}$ and setting the derivative to zero gives the following normal equations:

$$X^T X \underline{\beta} = X^T \underline{y}$$

- Now, if the matrix $X^T X$ is invertible (i.e. if X is of full rank), then the LSE of $\underline{\beta}$ is given by

$$\widehat{\underline{\beta}} = (X^T X)^{-1} X^T \underline{y}$$

Fitted Value

- The fitted value of response corresponding to regressor values $\underline{x} = (1, x_1, \cdots, x_p)$ is

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p$$

- Then $\underline{\hat{y}} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n)^T = X\underline{\widehat{\beta}} = X(X^TX)^{-1}X^T\underline{y} = H\underline{y}$, where, $H = X(X^TX)^{-1}X^T$ is called hat-matrix.
- The residuals are

$$\underline{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 - \hat{y}_1 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix} = \underline{y} - \underline{\hat{y}} = (I - H)\underline{y}$$

Properties of LSE of $\underline{\beta}$

- $\underline{\hat{\beta}}$ is a linear function of $\underline{y}$
- $\underline{\hat{\beta}}$ is unbiased estimator of $\underline{\beta}$. That is, $E(\underline{\hat{\beta}}) = \underline{\beta}$
- $Var(\underline{\hat{\beta}}) = \sigma^2(X^TX)^{-1}$
- $\underline{\hat{\beta}}$ is the BLUE of $\underline{\beta}$.

Estimation of Error Variance (σ^2)

- It can be shown that

$$E(SS_{Res}) = (n - p - 1)\sigma^2$$

- Hence, $\widehat{\sigma}^2 = \frac{SS_{Res}}{n-p-1} = MS_{Res}$ is an unbiased estimator of σ^2.
- Observed value of $\widehat{\sigma}^2 = \frac{SS_{Res}}{n-p-1}$ is called residual variance. It's positive square root is called residual standard error.
- Computationally efficient formula:

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = y^T y - \hat{\beta}_1^T X^T y,$$

Hypothesis Testing: Test for Significance of Regression

- Assumptions: ϵ_i's are i.i.d $N(0, \sigma^2)$ Rvs. Then $\underline{\epsilon} \sim N_n(\underline{0}, \sigma^2 I_n)$
- Want to test the hypothesis if there is a linear relationship between the response y and any of the regressor $x_1, \cdots, x_n$.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \text{ ag. } H_1 : \beta_j \neq 0 \text{ for atleast one } j$$

- Therefore, the test statistic is

$$F_0 = \frac{SS_{Reg}/p}{\hat{\sigma}^2} = \frac{SS_{Reg}/p}{SS_{Res}/(n-p-1)} \sim F_{p,n-p-1}, \text{ under } H_0.$$

- Reject H_0 iff $F_0 > F_{p,n-p-1;\alpha}$ (at level α).

Hypothesis Testing for individual regression coefficients: β_j

- Want to test:

$$H_0 : \beta_j = 0 \text{ ag. } H_1 : \beta_j \neq 0$$

- Therefore, the test statistic is

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{MS_{Res} C_{jj}}} \sim t_{n-p-1}, \text{ under } H_0,$$

  where C_{jj} is the diagonal element of $(X^T X)^{-1}$
- Reject H_0 iff $|t_0| > t_{n-p-1,\alpha/2}$; (at level α).

Test of contribution of a subset of the regressors

- Let us partition the problem as follows:

$$\underline{y} = \underline{X}\underline{\beta} + \underline{\epsilon} \implies \underline{y} = (\underline{X_1}, \underline{X_2}) \begin{pmatrix} \underline{\beta_1} \\ \underline{\beta_2} \end{pmatrix} + \underline{\epsilon}$$

- Want to test:

$$H_0 : \underline{\beta_2} = 0 \text{ ag. } H_1 : \underline{\beta_2} \neq 0$$

- Based on the full model ($\underline{y} = \underline{X}\underline{\beta} + \underline{\epsilon}$), $\underline{\hat{\beta}} = (X^T X)^{-1} X^T y$ and $SS_{Reg}(\underline{\beta})$ has p degrees of freedom.
- To find the contribution of $\underline{\beta_2}$, fit the model assuming $\underline{\beta_2} = 0$.
- The reduced model is $\underline{y} = \underline{X_1}\underline{\beta_1} + \underline{\epsilon}$, $\underline{\hat{\beta}_1} = (X_1^T X_1)^{-1} X_1^T y$ and $SS_{Reg}(\underline{\beta_1})$ has $p - r$ degrees of freedom. Where r denotes the number of components in $\underline{\beta_2}$

Test of contribution of a subset of the regressors

- $SS_{Reg}(\underline{\beta_2}|\underline{\beta_1}) = SS_{Reg}(\underline{\beta}) - SS_{Reg}(\underline{\beta_1})$ can be used as a measure of contribution of $\underline{\beta_2}$.
- Note that if $\underline{\beta_2}$ has significant contribution then $SS_{Reg}(\underline{\beta_2}|\underline{\beta_1})$ is large.
- Therefore, the test statistic is

$$F_0 = \frac{SS_{Reg}(\underline{\beta_2}|\underline{\beta_1})/r}{\hat{\sigma}^2} = \frac{SS_{Reg}(\underline{\beta_2}|\underline{\beta_1})/r}{SS_{Res}/(n-p-1)} \sim F_{r,n-p-1}, \text{ under } H_0.$$

- Reject H_0 iff $F_0 > F_{r,n-p-1;\alpha}$ (at level α).

Testing of general linear hypothesis

- Want to test:

$$H_0 : T\underline{\beta} = 0 \text{ ag. } H_1 : T\underline{\beta} \neq 0,$$

  where T is a $m \times (p+1)$ matrix of constants.
- Examples: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$
  - $H_0 : \beta_1 = \beta_3$ ag. $H_1 : \beta_1 \neq \beta_3$. Take, $T = \begin{bmatrix} 0 & 1 & 0 & -1 \end{bmatrix}$.
  - $H_0 : \beta_1 = \beta_3, \beta_2 = 0$ ag. $H_1 : \beta_1 \neq \beta_3$ or $\beta_2 \neq 0$. Take,

$$T = \begin{bmatrix} 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

  .
- The full model (FM) is $\underline{y} = \underline{X}\underline{\beta} + \underline{\epsilon}$.
- Under FM, $\underline{\hat{\beta}} = (X^T X)^{-1} X^T y$ and $SS_{Res}(FM) = y^T y - \hat{\beta}^T X^T y$ has $n - p - 1$ degrees of freedom.

Testing of general linear hypothesis

- Now assume that T has $r\,(\leq m)$ independent rows.
- Then $T\underline{\beta} = 0$ can be solved and r of the β_j's in FM can be written in terms of other $(p+1-r)$ β_j's.
- This lead to the reduced model (RM)

$$\underline{y} = \underline{Z}\underline{\gamma} + \underline{\epsilon},$$

  where Z is $n \times \overline{p+1-r}$ matrix.
- Under RM, $\hat{\underline{\gamma}} = (Z^T Z)^{-1} Z^T y$ and $SS_{Res}(RM) = y^T y - \hat{\gamma}^T Z^T y$ has $n-p-1+r$ degrees of freedom.
- $SS_{Res}(FM) \leq SS_{Res}(RM)$.

Testing of general linear hypothesis

- $SS_H = SS_{Res}(RM) - SS_{Res}(FM)$ with degrees of freedom $(n - p - 1 + r) - (n - p - 1) = r$.
- Therefore, the test statistic is

$$F_0 = \frac{SS_H/r}{SS_{Res}(FM)/(n-p-1)} \sim F_{r,n-p-1}, \text{ under } H_0.$$

- Reject H_0 iff $F_0 > F_{r,n-p-1;\alpha}$ (at level α).

Confidence Intervals (CIs)

- Confidence Interval of individual regression coefficient $\hat{\beta}_j$
- Pivot is

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{MS_{Res} C_{jj}}} \sim t_{n-p-1},$$

  where C_{jj} is the diagonal element of $(X^T X)^{-1}$ matrix.
- A $100(1-\alpha)\%$ CI for β_j is

$$\left[\hat{\beta}_j \pm t_{n-p-1,\alpha/2}\sqrt{MS_{Res} C_{jj}}\right].$$

CI for mean response

- Let $\underline{x_0} = \begin{pmatrix} 1 \\ x_{01} \\ \vdots \\ x_{0p} \end{pmatrix}$ be a value of the regressor vector.
- The mean response at $\underline{x_0}$ is $\underline{x_0}^T \underline{\beta}$.
- $\underline{x_0}^T \widehat{\beta} \sim N\left(\underline{x_0}^T \underline{\beta}, \sigma^2 \underline{x_0}^T \left(X^T X\right)^{-1} \underline{x_0}\right)$.
- Pivot is $\frac{\underline{\hat{y}_0} - \underline{x_0}^T \underline{\beta}}{\sqrt{MS_{Res}\, \underline{x_0}^T (X^T X)^{-1} \underline{x_0}}} \sim t_{n-p-1}$.
- A $100(1-\alpha)\%$ CI for mean response at $\underline{x_0}$ is

$$\left[\underline{\hat{y}_0} \pm t_{n-p-1,\alpha/2} \sqrt{MS_{Res}\, \underline{x_0}^T (X^T X)^{-1} \underline{x_0}}\right].$$

Prediction Interval

- Consider a level of regressor x_0.
- Let y_0 be the corresponding value of the response.
- We want a prediction interval for y_0.
- Let $\Psi = y_0 - \hat{y_0} = y_0 - \underline{x_0}^T \underline{\hat{\beta}} \sim N(0, \sigma^2(1 + \underline{x_0}^T (X'X)^{-1} \underline{x_0}))$.
- A pivot is $\frac{\Psi}{\sqrt{MS_{Res}(1+\underline{x_0}^T (X'X)^{-1} \underline{x_0})}} \sim t_{n-p-1}$.
- A $100(1-\alpha)\%$ prediction interval of y_0 is

$$[\hat{y_0} \mp t_{n-p-1;\frac{\alpha}{2}} \sqrt{MS_{Res}(1 + \underline{x_0}^T (X'X)^{-1} \underline{x_0})}].$$

Standardized Regression Coefficients

- Difficult to compare regression coefficients. The magnitude of β_j reflects the unit of measurement of regressor x_j.
- For example, $y = 5 + x_1 + 1000x_2$, where y is measured in liters, x_1 in milliliters, and x_2 in liters. Here, $\beta_2 >> \beta_1$. But the effects of both regressor on y are identical.
- Way-out is to standardized the regressors and response so that they become unit free.

Standardized Regression Coefficients

- A popular approach is as follows:
  - Define $W_{ij} = \frac{x_{ij} - \overline{x}_j}{\sqrt{S_{jj}}}$, $i = 1, 2, ..., n; j = 1, 2, ..., p$.
    $y_i^* = \frac{y_i - \overline{y}}{\sqrt{SS_T}}$, $i = 1, 2, ..., n$.
    Here $S_{jj} = \sum_{i=1}^{n} (x_{ij} - \overline{x_j})^2$, $j = 1, 2, ..., p$.
  - Clearly the mean $\overline{W_j} = 0$ and
    $\sqrt{\sum_{i=1}^{n} (W_{ij} - \overline{W}_j)^2} = \sqrt{\sum_{i=1}^{n} W_{ij}^2} = 1$
    $\overline{y}^* = 0$ and $(\sum_{i=1}^{n} (y_i^*)^2)^{\frac{1}{2}} = 1$
  - In terms of $y^*, W_1, ..., W_p$, the regression model becomes,

$$y^* = b_1 W_1 + b_2 W_2 + ... + b_p W_p + \epsilon$$

  - In matrix notation, $\underline{y}^* = W\underline{b} + \underline{\epsilon}$.
  - LSE, $\hat{\underline{b}} = (W'W)^{-1} W' \underline{y}^*$.

Standardized Regression Coefficients

- $$W'W = \begin{pmatrix} 1 & r_{12} & . & . & . & r_{1p} \\ r_{21} & 1 & . & . & . & r_{2p} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ r_{p1} & r_{p2} & . & . & . & 1 \end{pmatrix}_{p \times p} \qquad W'\underline{y}^* = \begin{pmatrix} r_{1y} \\ r_{2y} \\ . \\ . \\ . \\ r_{py} \end{pmatrix}_{p \times 1}$$

  where, $r_{ij} = \frac{\sum_{u=1}^{n}(x_{ui}-\bar{x}_i)(x_{uj}-\bar{x}_j)}{\sqrt{S_{jj}S_{ii}}} = \frac{S_{ij}}{\sqrt{S_{jj}S_{ii}}}$,

  $r_{iy} = \frac{\sum_{u=1}^{n}(x_{ui}-\bar{x}_i)(y_u-\bar{y})}{\sqrt{S_{ii}S_T}} = \frac{S_{jy}}{\sqrt{S_{ii}SS_T}}$

Adjusted R^2

- Coefficient of determination is given by

$$R^2 = 1 - \frac{SS_{Res}}{SS_T}.$$

- In multiple linear regression, adding a variable to a model can increase the value of R^2.
- To overcome this problem, we have adjusted R^2 defined by

$$R^2_{adj} = 1 - \frac{SS_{Res}/(n-p-1)}{SS_T/(n-1)}.$$

Model Adequacy Checking

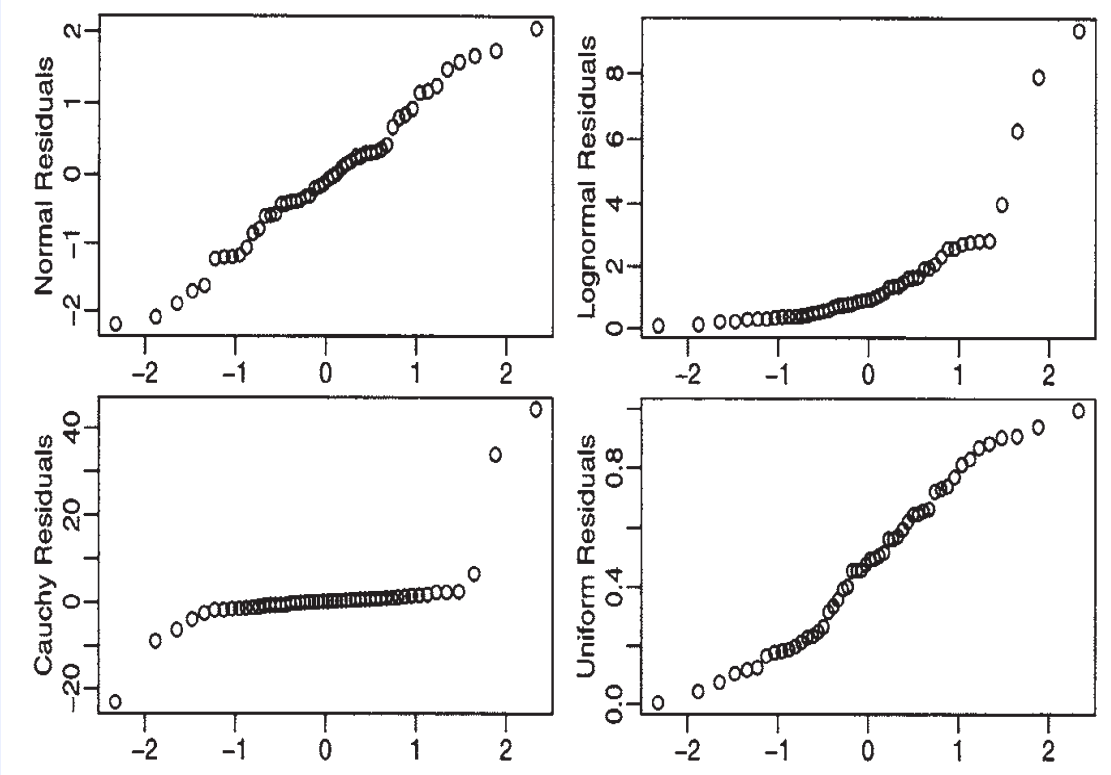
- Major assumptions
  - linear relationship
  - Error mean zero
  - Error variance is constant
  - Error are uncorrelated
  - Error are normally distributed and independent
- Gross violation of the assumptions may lead to a totally different model with opposite conclusions.
- We perform the checking using residuals.

Different Residuals

- We now define 3 types of residuals.
- Residual: $e_i = y_i - \hat{y}_i$ for all $i \implies \underline{e} = (I - H)\underline{y}$.
- Standardized residual:
  $d_i = \frac{e_i}{\sqrt{MS_{Res}}}$ for all $i \implies \underline{d} = \frac{1}{\sqrt{MS_{Res}}}(I - H)\underline{y}$.
- Studentized residual: $r_i = \frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}}$ for all i.

Residual Plots : Q–Q Plot

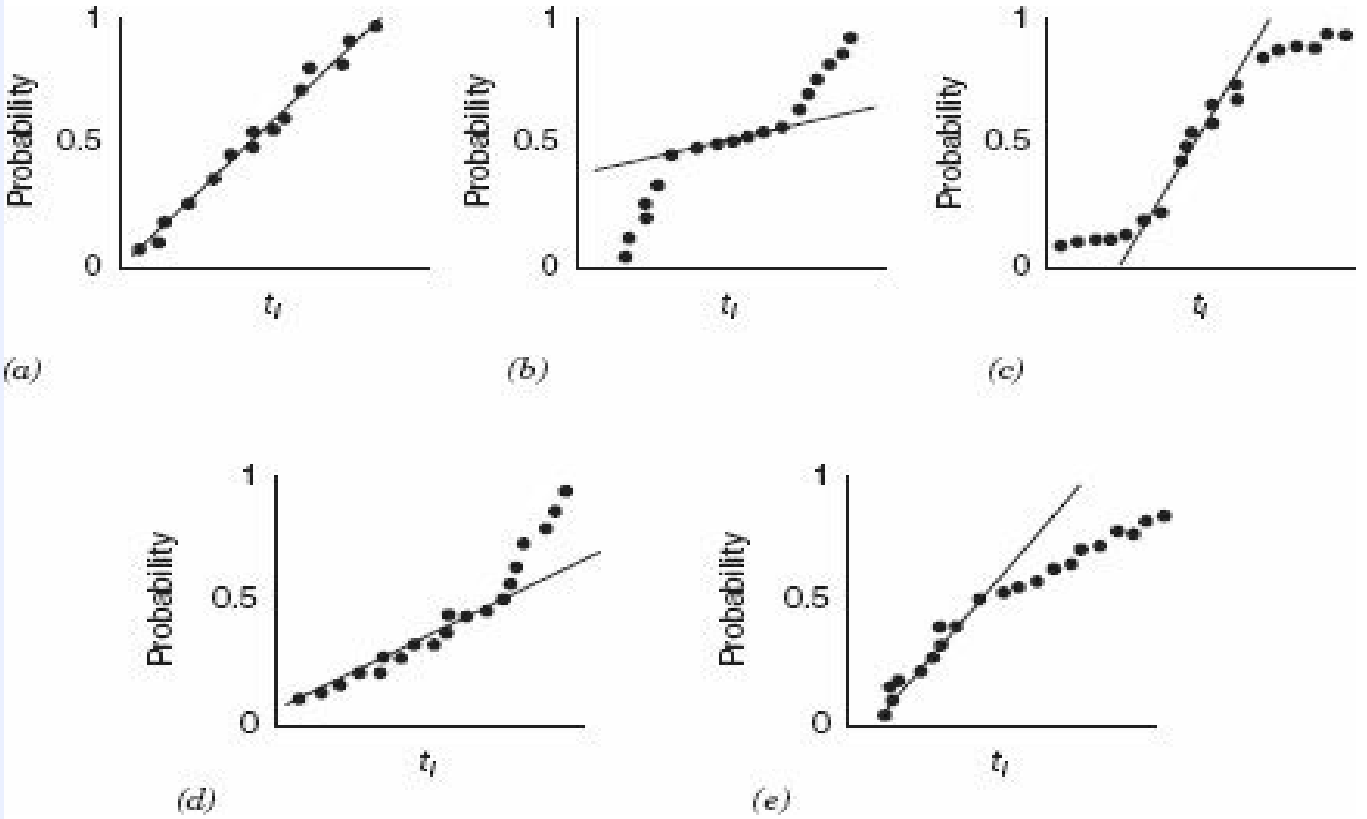
- Test for normality.
- The residuals can be assessed for normality using a Q–Q plot. This compares the residuals to "ideal" normal observations.
- We plot the quantiles corresponding to sorted residuals (e_i) against $\Phi^{-1}(\frac{i}{n+1})$ for $i = 1, \cdots, n$.



Source : Linear Models with R by Julian J. Faraway

Residual Plots : Q–Q Plot

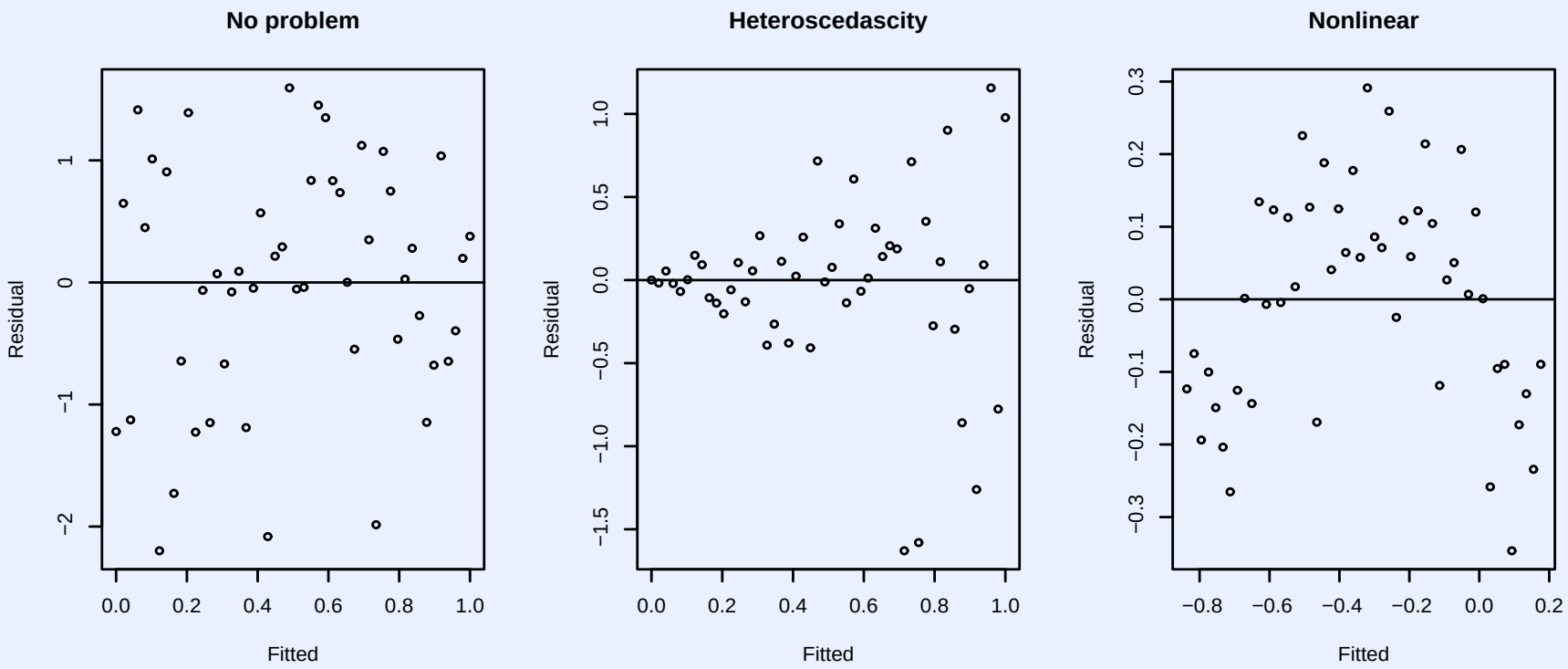
Figure 4.3 *Normal probability plots: (a) ideal; (b) light-tailed distribution; (c) heavy-tailed distribution; (d) positive skew; (e) negative skew.*



Source: Introduction to Linear Regression Analysis, by Montgomery, Peck, Vining; 2006.

Residual Plots: Plot of residual against Fitted values

- Test of constant variance and non-linear relation
- Plot $\hat{y}_i$ vs $e_i (or\ d_i\ or\ r_i)$



Residual Plots : Plot of residual against Regressors

- Plot x_{ij} vs e_i for all j
- In previous plot, replace 'fitted' by x_{ij} to find similar interpretation.
- Repeat it for all the regressors.

Partial Regression Plot

- Complete/correct marginal effect
- Let we want to study the marginal effect of x_j on y
- y is regressed on $x_1, ..., x_k$ except x_j.
- x_j is regressed on $x_1, ..., x_k$ except x_j.
- Plot y residual, $e_i(y|x_1, ..., x_{j-1}, x_{j+1}, ..., x_k)$ against x_j residual, $e_i(x_j|x_1, ..., x_{j-1}, x_{j+1}, ..., x_k)$
- For ideal scenario, the partial regression plot should show a linear relationship (straight line with non-zero slop).
- Curvilinear band: indicates higher order terms in x_j or it's transformation.
- Horizontal band: indicates no additional useful information in x_j

The PRESS Statistic

- PRESS: Prediction Error Sum of Squares
- Delete i^{th} observation. Fit the model on remaining $(n-1)$ observations and predict y_i.
- The corresponding predicted value is denoted by $\widehat{y}_{(-i)}$.
- The corresponding prediction error is $e_{(-i)} = y_i - \hat{y}_{(-i)}$. It is called PRESS residual.
- It can be shown that $e_{(-i)} = \frac{e_i}{1-h_{ii}}$.
- Large values of $e_{(-i)}$ implies potential influential observations.
- Large difference between e_i and $e_{(-i)}$ indicates an observation where model fit is quite well but a model built without that predicts poorly.

The PRESS Statistic

- $Var(e_{(-i)}) = \frac{\sigma^2}{1-h_{ii}}$.
- Standardized PRESS residual is

$$e_{(-i)}\sqrt{\frac{1-h_{ii}}{MS_{Res}}} = \frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}},$$

  which is same as the Studentized residual.
- $PRESS = \sum_{i=1}^{n} e_{(-i)}^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2$.
- PRESS is a measure of how well a regression model perform in predicting new observations.
- R^2 for prediction : $R^2_{prediction} = 1 - \frac{PRESS}{SS_T}$. It gives indication of the prediction capability of the regression model.
- Using PRESS, we may compare model.

Variable Selection

- Techniques:
  - All possible Regression.
  - Step-wise Type Procedures :
    - Forward Selection
    - Backward elimination
    - Step-wise Regression

Multicollinearity

- Near linear relationship among regressors.
- Effect of Multicollinearity – I :
  - Consider scaled response and regressor (length unit).
  - Consider $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$.
  - $\hat{\beta}_1 = \dfrac{r_{1y} - r_{12} r_{2y}}{1 - r_{12}^2}$, and $\hat{\beta}_2 = \dfrac{r_{2y} - r_{12} r_{1y}}{1 - r_{12}^2}$.
  - $Var(\hat{\beta}_j) = \dfrac{\sigma^2}{1 - r_{12}^2}$, $Cov(\hat{\beta}_1, \hat{\beta}_2) = \dfrac{-r_{12}\sigma^2}{1 - r_{12}^2}$.
  - Strong multicollinearity between x_1 and x_2 indicates the r_{12} will be large.
  - If $|r_{12}| \to 1$, $Var(\hat{\beta}_j) \to \infty$, and $|Cov(\hat{\beta}_1, \hat{\beta}_2)| \to \infty$.
  - The above large variances and covariances means different sample taken at the same x level could lead to widely different estimates of the model parameters.

Multicollinearity

- Effect of Multicollinearity – II:
  - $L_1^2 = (\widehat{\beta} - \beta)^T (\widehat{\beta} - \beta)$.
  - $E(L_1^2) = \sum_{j=1}^{p} Var(\widehat{\beta}_j) = \sigma^2 Tr(X'X)^{-1} = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$, where λ_j's are eigenvalues of $(X'X)$.
  - If $(X'X)$ is ill-conditioned then at least one λ_j will be small $\Rightarrow E(L_1^2)$ is big.
  - Therefore, we have $E(\underline{\widehat{\beta}}^T \underline{\widehat{\beta}}) = \underline{\beta}' \underline{\beta} + \sigma^2 Tr(X'X)^{-1}$, implies magnitude of $\underline{\widehat{\beta}}$ are large.

Multicollinearity Diagnostics

- Examination of correlation matrix $(X'X)$:
  - If x_i and x_j are nearly linearly dependent, then $|r_{ij}|$ should be close to 1.
  - However, this procedure is helpful to detect near linear dependence between a pair of regressors only.

Multicollinearity Diagnostics

- Variance Inflation Factors (VIFs):
  - $Var(\hat{\beta}_j) = \sigma^2 c_{jj}, C = (X'X)^{-1}$. It can be shown that $c_{jj} = (1 - R_j^2)^{-1} = \frac{1}{1-R_j^2}$, where R_j^2 is the coefficient of determination obtained when x_j is regressed on remaining $(k-1)$ regressors.
  - $VIF_j = \frac{1}{1-R_j^2}$: This measures the factor by which the variance of $\hat{\beta}_j$ inflated due to the near linear dependence.
  - Rule of thumb : If any of $VIF > 5$, the associated coefficient is estimated poorly due to multicollinearity.

Multicollinearity Diagnostics

- Eigen System Analysis of $(X'X)$:
  - The eigen values, $\lambda_1, \lambda_2, ..., \lambda_p$, can be used to see the extent of multicollinearity.
  - Small eigen values (one or more) $\Rightarrow$ multicollinearity.
  - Condition number, $k = \frac{\lambda_{max}}{\lambda_{min}}$.
  - Rule of thumb :
    - $k < 100 \rightarrow$ No serious problem with multicollinearity.
    - $100 \leq k < 1000 \rightarrow$ moderate to strong multicollinearity.
    - $k \geq 1000 \rightarrow$ severe multicollinearity.
  - Condition indices : $k_j = \frac{\lambda_{max}}{\lambda_j},\ j = 1, 2, ..., p$

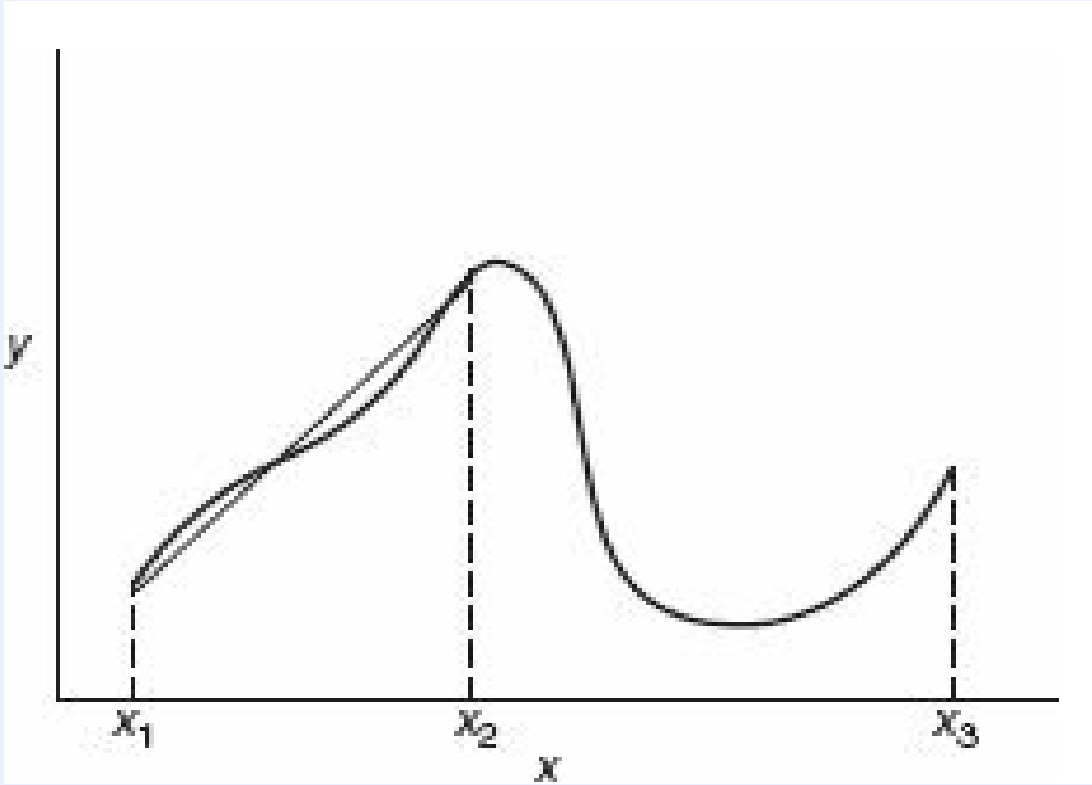
    The number of j's such that, $k_j \geq 1000 \rightarrow$ provide useful information on the number of near linear dependence.

Method for dealing with multicollinearity

- Source of multicollinearity:
  - Data collection method (ex: biased sample) $\rightarrow$ collecting more data.
  - Constraints in model or population (ex: family income (x_1) and household size (x_2)) $\rightarrow$ Model respecification
  - Model specification (ex: range of x is small, then adding x^2 in the model) $\rightarrow$ Model respecification
  - An overdefined model (ex: adding more regressors) $\rightarrow$ Model respecification, and other method of estimate like Ridge regression.

Considerations in the use of Regression

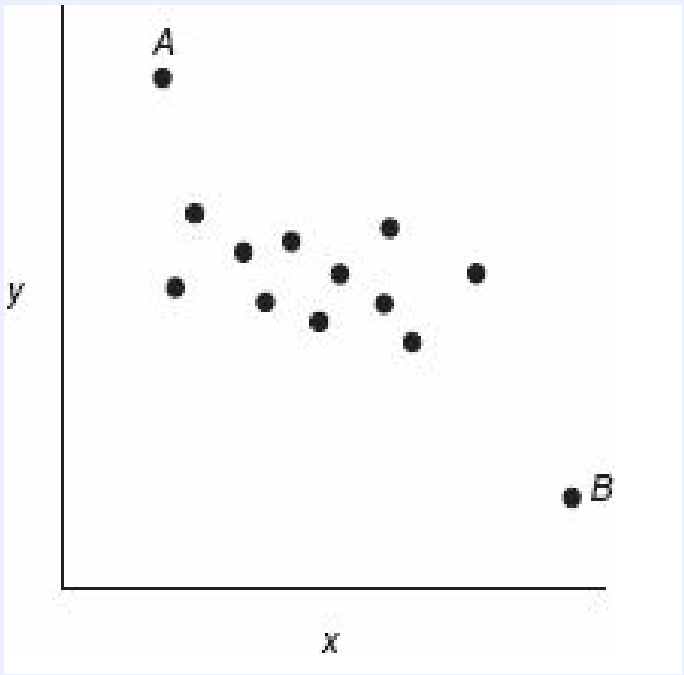
- Regression models are intended as interpolation equations over the range of the regressor variables used to fit the model.

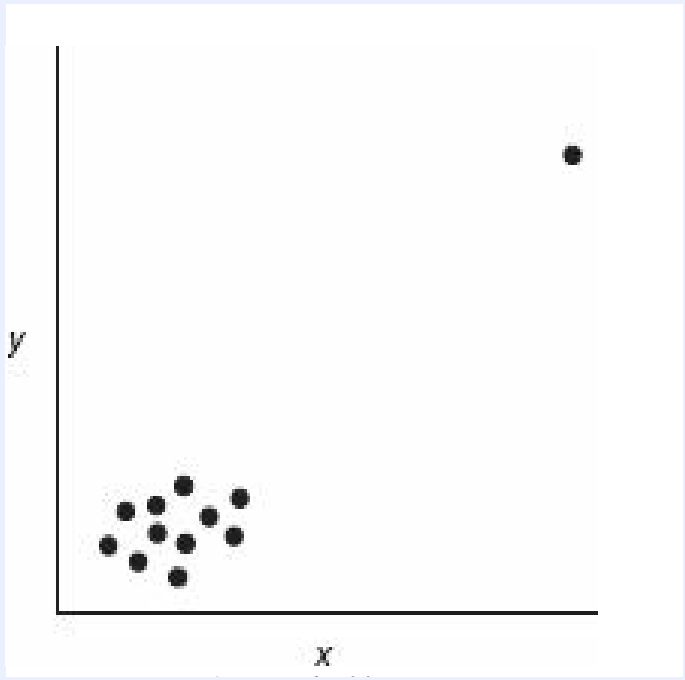


Source: Introduction to Linear Regression Analysis, by Montgomery, Peck, Vining; 2006.

Considerations in the use of Regression

- The deposition of the x-values plays an important role in the LS fit.

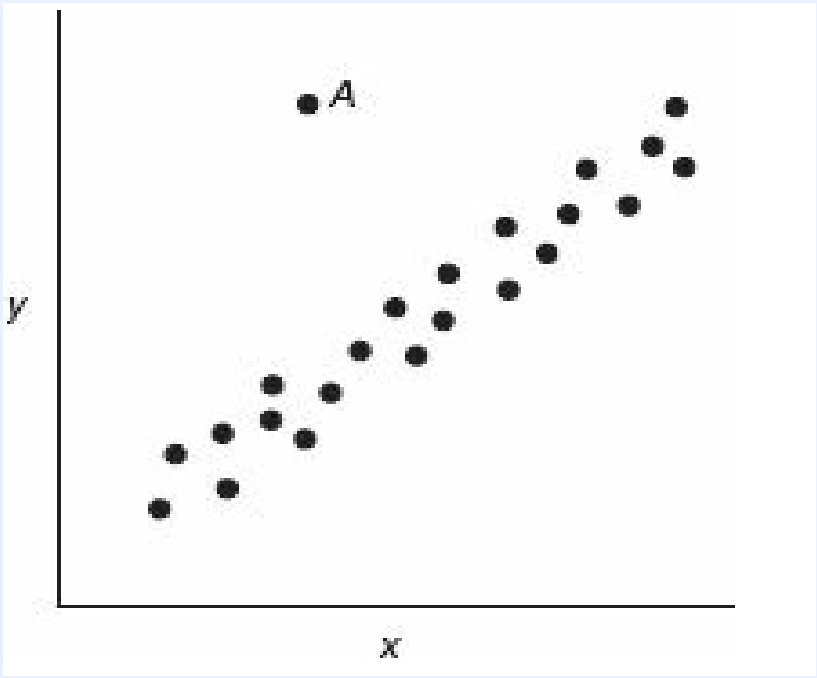




Source: Introduction to Linear Regression Analysis, by Montgomery, Peck, Vining; 2006.

Considerations in the use of Regression

- Outliers or bad values can seriously disturb the LS fit.



Source: Introduction to Linear Regression Analysis, by Montgomery, Peck, Vining; 2006.

Considerations in the use of Regression

- A regression analysis can only address the issues on correlation.
- It does not imply that the variables are relate in any causal sense.

Year	Number of Certified Mental Defectives per 10,000 of Estimated Population in the U.K (y)	Number of Radio Receiver Licenses Issued (Millions) in the U.K (x_1)	First Name of President of the U.S. (x_2)
1924	8	1.350	Calvin
1925	8	1.960	Calvin
1926	9	2.270	Calvin
1927	10	2.483	Calvin
1928	11	2.730	Calvin
1929	11	3.091	Calvin
1930	12	3.647	Herbert
1931	16	4.620	Herbert
1932	18	5.497	Herbert
1933	19	6.260	Herbert
1934	20	7.012	Franklin
1935	21	7.618	Franklin
1936	22	8.131	Franklin
1937	23	8.593	Franklin

Source: Kendall and Yule [1950] and Tufte [1974].

- The fitted regression equation relating y to x_1 is

$$\widehat{y} = 4.58 + 2.20x_1.$$

- $R^2 = 0.9842$.
- For testing $H_0 : \beta_1 = 0$, the p-value is 3.58×10^{-12}.

How to attack the data analysis/model fitting:

- Understand the research question(s). Understand the data you have.
  1. How the data was collected?
  2. What type of the study design used: *Randomized or Observational; Prospective or Retrospective etc.*
  3. Can you make connection with the primary research question and the data? Is the research question feasible based on the data you have?
  4. Are there secondary research questions?
  5. What are the potential source of bias? Sample (data) may not be a representative of the target (source) population.
  6. Are there any confounders?
  7. Are the number of observations/individuals in the data sufficient?

How to attack the data analysis/model fitting:

- Do the scatter plot(s): response vs. input variable(s).
- Fit the regression model(s) (or other type of model(s)).
- Interpret the output from the fitted models:
  1. Are all the results expected? Whether the results go well with existing domain (basic science) knowledge?
  2. If not, what are the reasons behind the aberration from the expected results.
- Check the diagnostics for model assumptions. If you find problem, go back and correct (if you can) the chosen model; or, take decision about the outliers/influential points.