Statistical Regression Analysis
Introduction to Multiple Linear Regression

Ernest Fokoué

School of Mathematical Sciences
Rochester Institute of Technology
Rochester, New York, USA

Statistical Regression Analysis
STAT 741-Autumn Semester 2018

January 8, 2021

Introduction to Multiple Linear Regression (MLR)

Upon completing this session, you will have learned

1. *Basic Aspects of Multiple Linear Regression*
   - *General Formulation of MLR*
   - *Tools for Analyzing the MLR model*
2. *Estimation, Inference and Prediction for MLR*
   - *Estimation of Coefficients by adaptation of MLR*
   - *Appeal of MLR*
   - *Properties of the OLS estimates*
   - *Inference and Prediction with MLR*
3. *Fitting Multiple Linear Regression Models in R*
   - *Real data example*
   - *Simulated data example*
4. *Exercises for exploration*

Basic Formulation of MLR

Given $\{(\mathbf{x}_i, \mathrm{y}_i),\ i = 1, \cdots, n\}$. *where* $\mathbf{x}_i^\top = (\mathrm{x}_{i1}, \cdots, \mathrm{x}_{ip})$ *and* $\mathrm{y}_i \in \mathbb{R}$. *Assuming the multiple linear regression (MLR) model*

$$Y_i = \beta_0 + \beta_1 \mathrm{x}_{i1} + \beta_2 \mathrm{x}_{i2} + \cdots + \beta_p \mathrm{x}_{ip} + \epsilon_i, \tag{1}$$

When Equation (1) *is applied to the whole training set, we get*

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & \mathrm{x}_{11} & \mathrm{x}_{12} & \cdots & \mathrm{x}_{1j} & \cdots & \mathrm{x}_{1p} \\ 1 & \mathrm{x}_{21} & \mathrm{x}_{22} & \cdots & \mathrm{x}_{2j} & \cdots & \mathrm{x}_{2p} \\ \vdots & \vdots & \vdots & \ldots & \ddots & \ldots & \vdots \\ 1 & \mathrm{x}_{i1} & \mathrm{x}_{i2} & \cdots & \mathrm{x}_{ij} & \cdots & \mathrm{x}_{ip} \\ \vdots & \vdots & \vdots & \ldots & \ddots & \ldots & \vdots \\ 1 & \mathrm{x}_{n1} & \mathrm{x}_{n2} & \cdots & \mathrm{x}_{nj} & \cdots & \mathrm{x}_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{bmatrix}$$

As a result, using $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)$, $\boldsymbol{Y}^\top = (Y_1, Y_2, \cdots, Y_n)$ *and* $\boldsymbol{\epsilon}^\top = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)$, *we can write*

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Basic Formulation of MLR

Given $\{(\mathbf{x}_i, \mathrm{y}_i),\ i = 1, \cdots, n\}$. *where* $\mathbf{x}_i^\top = (\mathrm{x}_{i1}, \cdots, \mathrm{x}_{ip})$ *and* $\mathrm{y}_i \in \mathbb{R}$. *The multiple linear regression model of Equation* (1) *has the matrix form*

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{X}$ *is the* $n \times (p+1)$ *design matrix defined by*

$$\mathbf{X} = \begin{bmatrix} 1 & \mathrm{x}_{11} & \mathrm{x}_{12} & \cdots & \mathrm{x}_{1j} & \cdots & \mathrm{x}_{1p} \\ 1 & \mathrm{x}_{21} & \mathrm{x}_{22} & \cdots & \mathrm{x}_{2j} & \cdots & \mathrm{x}_{2p} \\ \vdots & \vdots & \vdots & \cdots & \ddots & \cdots & \vdots \\ 1 & \mathrm{x}_{i1} & \mathrm{x}_{i2} & \cdots & \mathrm{x}_{ij} & \cdots & \mathrm{x}_{ip} \\ \vdots & \vdots & \vdots & \cdots & \ddots & \cdots & \vdots \\ 1 & \mathrm{x}_{n1} & \mathrm{x}_{n2} & \cdots & \mathrm{x}_{nj} & \cdots & \mathrm{x}_{np} \end{bmatrix} \in \mathbb{R}^{n\times(p+1)}.$$

and $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p) \in \mathbb{R}^{p+1}$, $\boldsymbol{Y}^\top = (Y_1, Y_2, \cdots, Y_n) \in \mathbb{R}^n$, $\boldsymbol{\epsilon}^\top = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n) \in \mathbb{R}^n$. *Note: The matrix* $\boldsymbol{X}$ *will (as expected) play a very important role in regression.*

The Ubiquitous Multivariate Gaussian Distribution

Let Y be an n-dimensional random vector. If Y is known (or assumed) to follow a multivariate Gaussian (normal) distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and variance-covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$, then we can write

- *Distributional declaration (denotation)*

$$Y \sim \texttt{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \textit{or} \quad Y \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- *Probability Density Function of Y*

$$p(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \right\} \tag{3}$$

  - $|\boldsymbol{\Sigma}| = \texttt{det}(\boldsymbol{\Sigma})$ *is the determinant of* $\boldsymbol{\Sigma}$
  - $\boldsymbol{\Sigma}^{-1}$ *is the inverse of* $\boldsymbol{\Sigma}$
  - $\boldsymbol{y}$ *is a realization of* Y

Ordinary Least Squares Estimation for MLR

Let $\mathbf{x}_i^\top = (\mathrm{x}_{i1}, \mathrm{x}_{i2}, \cdots, \mathrm{x}_{ip})$ *an input vector, and* $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)$ *the vector of regression coefficients. Then*

- *The error (noise) at point* i *is*

$$\epsilon_i = \mathrm{y}_i - \tilde{\mathbf{x}}_i^\top \boldsymbol{\beta} = \mathrm{y}_i - \sum_{j=0}^{p} \beta_j \mathrm{x}_{ij}$$

- *The error vector for the whole set of points is*

$$\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$$

- *The Sum of Squares of Errors (SSE) is given by*

$$SSE(\boldsymbol{\beta}) = \sum_{i=1}^{n} (\mathrm{y}_i - \tilde{\mathbf{x}}_i^\top \boldsymbol{\beta})^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{4}$$

Aspects of MLR

Assuming the MLR model with constant noise (error) variance, we have

- $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}] = \boldsymbol{X}\beta$
- $\mathbb{V}[\boldsymbol{Y}|\boldsymbol{X}] = \sigma^2 \boldsymbol{I}_n$

If $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, *we can write*

- *Distribution of error vector*

$$\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

- *Conditional distribution of response vector*

$$\boldsymbol{Y}|\boldsymbol{X} \sim N_n(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_n)$$

Appeal of multiple linear regression (MLR)

Important aspects of the classical multiple linear regression (MLR) model as presented in Clarke, Fokoué, Zhang (2009), Chap 2, page 54, are presented here. Consider the generic MLR model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

where the $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$*, independent of* $X_1, X_2, \cdots, X_p$*. The benefits of MLR are well known and include:*

- *MLR is interpretable - the effect of each predictor variable is captured by a single coefficient.*
- *Theory supports inference for the regression coefficients* β_j*'s.*
- *Prediction is easy and straightforward.*
- *Simple interactions between* X_i *and* X_j *are easy to include.*
- *Transformations of the* X_j*'s are easy to include and dummy variables allow the use of categorical information.*
- *Computation is fast.*

Least Squares Estimation for MLR

The OLS estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is the minimizer of SSE($\boldsymbol{\beta}$), namely

$$\hat{\boldsymbol{\beta}}^{(\mathtt{OLS})} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min}\,\{SSE(\boldsymbol{\beta})\} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min}\left\{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^{\top}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\right\},$$

which is obtained by taking the partial derivative

$$\frac{\partial SSE(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^{\top}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^{\top}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}),$$

then solving

$$\frac{\partial SSE(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \quad \textit{i.e.} \quad -2\boldsymbol{X}^{\top}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) = 0,$$

which yields the normal equations

$$\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^{\top}\boldsymbol{Y},$$

from which one gets

$$\hat{\boldsymbol{\beta}}^{(\mathtt{OLS})} = \hat{\boldsymbol{\beta}} = (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\boldsymbol{Y}. \tag{5}$$

Properties of the OLS Estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$

Given the least squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$,

- *The fitted value at point i is*

$$\hat{Y}_i = \tilde{\mathbf{x}}_i^\top \hat{\boldsymbol{\beta}}$$

- *The vector $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \cdots, \hat{Y}_n)^\top$ of fitted values is*

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

- *It turns out that $\hat{\boldsymbol{Y}}$ can be written as*

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$$

where the matrix

$$\boldsymbol{H} = \boldsymbol{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$$

is referred to as the hat matrix or projection.

Properties of the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$

The hat matrix $\boldsymbol{H}$ *plays a great role in regression analysis*

- $\boldsymbol{H}$ *is symmetric, i.e.* $\boldsymbol{H}^\top = \boldsymbol{H}$

$$\boldsymbol{H}^\top = (\boldsymbol{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)^\top = \boldsymbol{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \boldsymbol{H}$$

- $\boldsymbol{H}$ *is idempotent, which means that* $\boldsymbol{H}^m = \boldsymbol{H}$, $m \geq 1$. *Indeed, it is easy to verify that*

$$\boldsymbol{H}^2 = \boldsymbol{H}\boldsymbol{H} = \boldsymbol{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \boldsymbol{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \boldsymbol{H}$$

and then use induction principle to confirm that

$$\boldsymbol{H}_{m+1} = \boldsymbol{H}^m\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}^2 = \boldsymbol{H}$$

- $\boldsymbol{I}_n - \boldsymbol{H}$ *is also symmetric, so that* $(\boldsymbol{I}_n - \boldsymbol{H})^\top = \boldsymbol{I}_n - \boldsymbol{H}$
- $\boldsymbol{I}_n - \boldsymbol{H}$ *is also idempotent, i.e.* $(\boldsymbol{I}_n - \boldsymbol{H})^m = \boldsymbol{I}_n - \boldsymbol{H}$, $m \geq 1$

Important Matrices for Regression Analysis

- *The hat matrix* $\boldsymbol{H} \in \mathbb{R}^{n \times n}$ *is indeed* $n \times n$ *matrix whose entries are*

$$\boldsymbol{H} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \ddots & \cdots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix}$$

- *An important result about the hat matrix* $\boldsymbol{H}$ *is the fact*

$$\sum_{j=1}^{n} h_{ij} = 1$$

where

$$h_{ij} = \tilde{\mathbf{x}}_i^{\top} (\mathbf{X}^{\top}\mathbf{X})^{-1} \tilde{\mathbf{x}}_j$$

It can also be shown that $h_{ij} \geq 0$. *Also*

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j.$$

Degrees of freedom in linear regression analysis

- *The diagonal elements h_{ii} of $\boldsymbol{H}$ are crucial in residual analysis, precious because*

$$h_{ii} = \tilde{\mathbf{x}}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \tilde{\mathbf{x}}_i$$

  is the leverage of the ith observation.
- *It can be shown that*

$$\frac{1}{n} \leq h_{ii} < 1, \quad i = 1, 2, \cdots, n.$$

- *The trace of $\boldsymbol{H}$, the sum of the diagonal elements of $\boldsymbol{H}$, plays a vital role in inference for regression. Indeed, it can be shown that*

$$trace(\boldsymbol{H}) = \sum_{i=1}^{n} h_{ii} = h_{11} + h_{22} + \cdots + h_{nn} = p + 1$$

  Indeed, $df(RSS) = n - trace(\boldsymbol{H}) = n - (p + 1)$ is the number of degrees of freedom in the residual sum of squares (RSS), which appears in nearly every operations in regression.

Important Matrices for Regression Analysis

- *Residual for the ith observation*

$$e_i = Y_i - \hat{Y}_i = Y_i - \tilde{\mathbf{x}}_i^\top \boldsymbol{\beta}$$

- *Residual vector*

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}$$

- *Residual Sum of Squares in matrix form*

$$\begin{aligned} RSS(\hat{\boldsymbol{\beta}}) = SSE(\hat{\boldsymbol{\beta}}) &= \sum_{i=1}^{n} (\mathrm{y}_i - \tilde{\mathbf{x}}_i^\top \boldsymbol{\beta})^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^\top (\boldsymbol{Y} - \hat{\boldsymbol{Y}}) \\ &= (\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y})^\top (\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y}) \\ &= \boldsymbol{Y}^\top (\boldsymbol{I}_n - \boldsymbol{H})^\top (\boldsymbol{I}_n - \boldsymbol{H}) \boldsymbol{Y} \\ &= \boldsymbol{Y}^\top (\boldsymbol{I}_n - \boldsymbol{H})^2 \boldsymbol{Y} \\ &= \boldsymbol{Y}^\top (\boldsymbol{I}_n - \boldsymbol{H}) \boldsymbol{Y} \end{aligned}$$

- *Indeed* $SSE(\hat{\boldsymbol{\beta}}) = \boldsymbol{e}^\top \boldsymbol{e}$

Least Squares Estimation of the noise variance σ^2

- The usual estimator $\widehat{\sigma^2}$ of the noise variance σ^2 is given by

$$\widehat{\sigma^2} = \frac{SSE(\hat{\beta})}{n-p-1} = \frac{\boldsymbol{Y}^\top(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}}{n-p-1}$$

- It is also known that

$$\frac{(n-p-1)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-p-1}$$

- Which means that

$$SSE(\hat{\beta}) \sim \sigma^2 \chi^2_{n-p-1}$$

Matrix Form of Regression Analysis Components

Let $\boldsymbol{I}_n \in \mathbb{R}^{n\times n}$ *and* $\boldsymbol{J}_n \in \mathbb{R}^{n\times n}$ *be two* $n \times n$ *matrices defined as*

$$\boldsymbol{J}_n = \frac{1}{n}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \ddots & \cdots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{I}_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

The vector $\mathbf{1}_n^\top = (1, 1, \cdots, 1)$ *be an n-dimensional vector of ones.*

- *Residual Sum of Squares (RSS)*

$$RSS = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^\top(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = (\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y})^\top(\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y}) = \boldsymbol{Y}^\top(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}$$

- *Total variation in Y*

$$SST = (\boldsymbol{Y} - \boldsymbol{J}_n\boldsymbol{Y})^\top(\boldsymbol{Y} - \boldsymbol{J}_n\boldsymbol{Y}) = \boldsymbol{Y}^\top(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}$$

- *Sum of Squares of Regression*

$$SSR = SST - RSS = \boldsymbol{Y}^\top(\boldsymbol{H} - \boldsymbol{J}_n)\boldsymbol{Y}$$

Properties of the OLS Estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$

Recall that the OLS Estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}^{(\texttt{OLS})} = \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{Y}.$$

$\hat{\boldsymbol{\beta}}$ has the following nice properties:

- *$\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$, i.e. $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$,*
- *The variance-covariance matrix of $\boldsymbol{\beta}$ is*

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

- *If $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, then $\hat{\boldsymbol{\beta}}$ is normally distributed, specifically*

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$$

so that

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 ((\mathbf{X}^\top \mathbf{X})^{-1})_{jj})$$

Least Squares Estimation of the average response

- $\hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ *is the estimator of the average response vector* $\mathbb{E}[\boldsymbol{Y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$.
- *The estimator* $\hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ *is unbiased, since we have*

$$\mathbb{E}[\hat{\boldsymbol{Y}}] = \mathbf{X}\boldsymbol{\beta} = \mathbb{E}[\boldsymbol{Y}|\mathbf{X}]$$

- *The variance of* $\hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ *is* $\mathbb{V}[\hat{\boldsymbol{Y}}|\boldsymbol{X}] = \sigma^2\boldsymbol{H}$.

$$\begin{aligned}\mathbb{V}[\hat{\boldsymbol{Y}}|\boldsymbol{X}] &= \mathbb{V}[\mathbf{X}\hat{\boldsymbol{\beta}}] = \mathbf{X}\mathbb{V}[\hat{\boldsymbol{\beta}}]\boldsymbol{X}^\top = \boldsymbol{X}\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{X}^\top \\ &= \sigma^2\boldsymbol{X}(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{X}^\top = \sigma^2\boldsymbol{H}\end{aligned}$$

- *For the* i*th observation, it is*

$$\mathbb{V}[\hat{Y}_i|\mathbf{x}_i] = \sigma^2 h_{ii}$$

- *For a single new observation with* $\hat{Y}_0 = \tilde{\mathbf{x}}_0^\top\hat{\boldsymbol{\beta}}$,

$$\mathbb{V}[\hat{Y}_0|\mathbf{x}_0] = \sigma^2\tilde{\mathbf{x}}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\tilde{\mathbf{x}}_0$$

Distributional Results for MLR

Assuming that $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, *which the same as* $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, *we have the following distributional results crucially needed in inference on parameters and the construction of confidence and prediction intervals*

- *The vector* $\hat{\boldsymbol{\beta}}$ *has a normal distribution*

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$$

- *The residual sum of squares has a Chi-squared distribution with* $n - p - 1$ *degrees of freedom*

$$SSE(\hat{\boldsymbol{\beta}}) \sim \sigma^2 \chi^2_{n-p-1}$$

- *The estimated average response* $\hat{Y}_0 = \tilde{\mathbf{x}}_0^\top \hat{\boldsymbol{\beta}}$ *has a normal distribution*

$$\hat{Y}_0 = \tilde{\mathbf{x}}_0^\top \hat{\boldsymbol{\beta}} \sim N(\tilde{\mathbf{x}}_0^\top \boldsymbol{\beta}, \sigma^2 \tilde{\mathbf{x}}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \tilde{\mathbf{x}}_0)$$

- *We also have* $\hat{\boldsymbol{Y}} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{H})$.

Additional Distributional Results for MLR

The following facts hold true as a result of the basic assumptions

- $\mathbb{E}[\boldsymbol{e}|\mathbf{X}] = \mathbf{0}$ *and* $\mathbb{V}[\boldsymbol{e}|\mathbf{X}] = \sigma^2(\boldsymbol{I}_n - \boldsymbol{H})$
- *Thanks to the fact that* $cov(\hat{\boldsymbol{\beta}}, \boldsymbol{e}) = \mathbf{0}$*, we know that* $\hat{\boldsymbol{\beta}}$ *and* $\widehat{\sigma^2}$ *are independent.*
- *The estimated response are uncorrelated with the residuals, i.e.* $cov(\hat{\boldsymbol{Y}}, \boldsymbol{e}) = \mathbf{0}$

Assuming that $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$*, which the same as* $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$*,*

- *The residual vector* $\boldsymbol{e}$ *has a normal distribution*

$$\boldsymbol{e}|\mathbf{X} \sim N_n(\mathbf{0}, \sigma^2(\boldsymbol{I}_n - \boldsymbol{H}))$$

- *The distribution of the noise for observation* i *is normal*

$$e_i|\mathbf{x}_i \sim N(0, \sigma^2(1 - h_{ii}))$$

Main Test of Significance with MLR

- *Defining the test: Indeed the most fundamental thing to assess upon completing the fitting of a regression model, is to check if that fitted model is significant. In this case, the basic test, consist of finding out if at least one of the p predictor variables posited are significantly related to the response.*

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_j = \cdots = \beta_p = 0$$
$$H_a : \textbf{At least one } \beta_j \textbf{ is nonzero}$$

In other words, if even only one of the p variables is significant, the whole regression will be declared significant.

- *Test statistic*

$$F = \frac{(SST - SSE)/p}{SSE/(n-p-1)} \sim F_{p,n-p-1}$$

- *Significance of regression vs goodness of fit: The fact of getting significance does NOT necessarily mean that the posited model perfectly fits the data. Indeed, it could be that there are many other factors.*

Under the Gaussian noise assumption, with σ^2 is unknown, and n small,

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\mathtt{ese}(\hat{\beta}_j)} \sim t_{n-p-1}$$

To perform the significance test for the jth regression coefficient, i.e.,

$$H_0 : \beta_j = 0 \quad \textit{versus} \quad H_a : \beta_j \neq 0$$

simply use the test statistic

$$t_j = \frac{\hat{\beta}_j}{\mathtt{ese}(\hat{\beta}_j)}$$

where

$$\mathtt{ese}(\hat{\beta}_j) = \mathtt{s}\sqrt{(\mathbf{X}^\top \mathbf{X})^{-1})_{jj}}$$

with

$$\mathtt{s} = \sqrt{\frac{SSE}{n-p-1}}.$$

A $100(1-\alpha)\%$ *confidence interval for* β_j *is simply given by*

$$\hat{\beta}_j \pm t_{n-p-1,\alpha/2} \times \mathsf{ese}(\hat{\beta}_j)$$

where

$$t_{n-p-1,\alpha/2} = F^{-1}_{t_{n-p-1}}\left(1-\frac{\alpha}{2}\right)$$

so that

$$\Pr[T_{n-p-1} \leq t_{n-p-1}, \alpha/2] = 1-\frac{\alpha}{2}$$

Finally, the so-called multiple coefficient of determination is given by

$$R^2 = 1-\frac{SSE}{SST}$$

and the value of the F statistic is given by

$$F = \frac{(SST-SSE)/p}{SSE/(n-p-1)}$$

Estimated Average Response with MLR

Given a new point $\mathbf{x}_0^\top = (\mathrm{x}_{01}, \cdots, \mathrm{x}_{0p})$ *for which an estimated average response value is desired, form* $\tilde{\mathbf{x}}_0^\top = (1, \mathrm{x}_{01}, \mathrm{x}_{02}, \cdots, \mathrm{x}_{0p})$*, then simply compute*

$$\hat{\mu}(\mathbf{x}_0) = \widehat{\mathbb{E}[Y|\mathbf{x}_0]} = \sum_{j=0}^{p} \hat{\beta}_j \mathrm{x}_{0j} = \tilde{\mathbf{x}}_0^\top \hat{\boldsymbol{\beta}}$$

A corresponding $100(1-\alpha)\%$ *confidence interval for* $\mu(\mathbf{x}_0) = \mathbb{E}[Y|X = \mathbf{x}_0]$ *is simply given by*

$$\tilde{\mathbf{x}}_0^\top \hat{\boldsymbol{\beta}} \pm t_{n-p-1,\alpha/2} \times \mathsf{ese}(\hat{\mu}(\mathbf{x}_0))$$

where

$$\mathsf{ese}(\hat{\mu}(\mathbf{x}_0)) = \mathrm{s}\sqrt{\tilde{\mathbf{x}}_0(\mathbf{X}^\top \mathbf{X})^{-1}\tilde{\mathbf{x}}_0}$$

Prediction with MLR

Given a new point $\mathbf{x}_0 \in \mathbb{R}$ *for which a single response value is desired, form* $\tilde{\mathbf{x}}_0^\top = (1, \mathrm{x}_{01}, \mathrm{x}_{02}, \cdots, \mathrm{x}_{0p})$*, then simply compute*

$$\hat{Y}(\mathbf{x}_0) = \widehat{[Y|\mathbf{x}_0]} = \sum_{j=0}^{p} \hat{\beta}_j \mathrm{x}_{0j} = \tilde{\mathbf{x}}_0^\top \hat{\boldsymbol{\beta}}$$

A corresponding $100(1-\alpha)\%$ *prediction interval for* $[Y|\mathbf{x}_0]$ *is given by*

$$\tilde{\mathbf{x}}_0^\top \hat{\boldsymbol{\beta}} \pm t_{n-p-1,\alpha/2} \times \mathtt{pese}(\hat{Y}(\mathbf{x}_0))$$

where

$$\mathtt{pese}(\hat{Y}(\mathbf{x}_0)) = \mathrm{s}\sqrt{1 + \tilde{\mathbf{x}}_0(\mathbf{X}^\top \mathbf{X})^{-1}\tilde{\mathbf{x}}_0}$$

Performing MLR in R

Let consider our example of the motor cars datasets and then perform MLR. The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). Here we will keep only 5 of the predictor variables.

mpg	*Miles/(US) gallon*
disp	*Displacement (cu.in.)*
hp	*Gross Horsepower*
drat	*Rear axle ratio*
wt	*Weight (lb/1000)*
qsec	*1/4 mile time.*

The response variable in this case is mpg.

Performing MLR in R

We seek to build the multiple linear regression model

$$\texttt{mpg}_i = \beta_0 + \beta_1 \texttt{disp}_i + \beta_2 \texttt{hp}_i + \beta_3 \texttt{drat}_i + \beta_4 \texttt{wt}_i + \beta_5 \texttt{qsec}_i + \epsilon_i$$

with added assumption that $\epsilon_i \sim N(0, \sigma^2)$. *In R, this model fitting operation is done using*

```
lm.mpg <- lm(mpg~disp+hp+drat+wt+qsec, data=mycars)
```

Noticing that the predictor variables invoked (used) are ALL the predictor variables in the data being used, the above command can be simplified by simply using

```
lm.mpg <- lm(mpg~., data=mycars)
```

The dot $(\cdot)$ *accounts for all the columns of the data set. This way of fitting the MLR can be very handy when the number of predictor variables is large.*

Performing MLR in R

It is excellent practice to always take a peek at the data prior to doing anything on it. In R, we can see the first 6 rows using head() and the last 6 using tail().

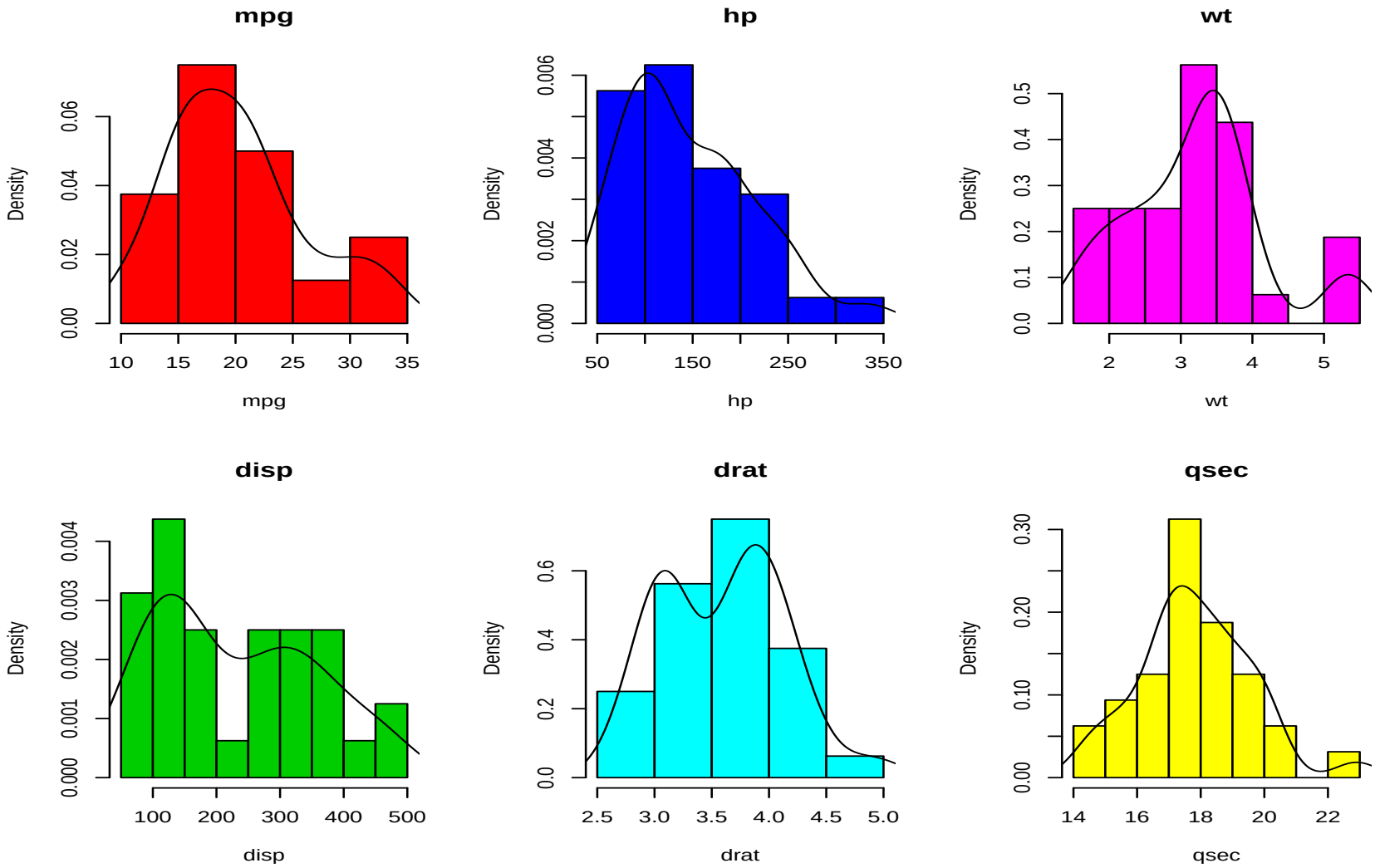
```
head(mycars)
```

	mpg	disp	hp	drat	wt	qsec
Mazda RX4	21.0	160	110	3.90	2.620	16.46
Mazda RX4 Wag	21.0	160	110	3.90	2.875	17.02
Datsun 710	22.8	108	93	3.85	2.320	18.61
Hornet 4 Drive	21.4	258	110	3.08	3.215	19.44
Hornet Sportabout	18.7	360	175	3.15	3.440	17.02
Valiant	18.1	225	105	2.76	3.460	20.22

It is clear from the above that the variables are measured on somewhat different scales. That difference of measurement scale can potential be a source of problems. One way around it is to standardized or cubitized the variables.

Performing MLR in R

Check the summary of each variable, both numerically and graphically.



Performing MLR in R

To print/display the correlation matrix in R, simply use

```
print(round(cor(mycars),2))
```

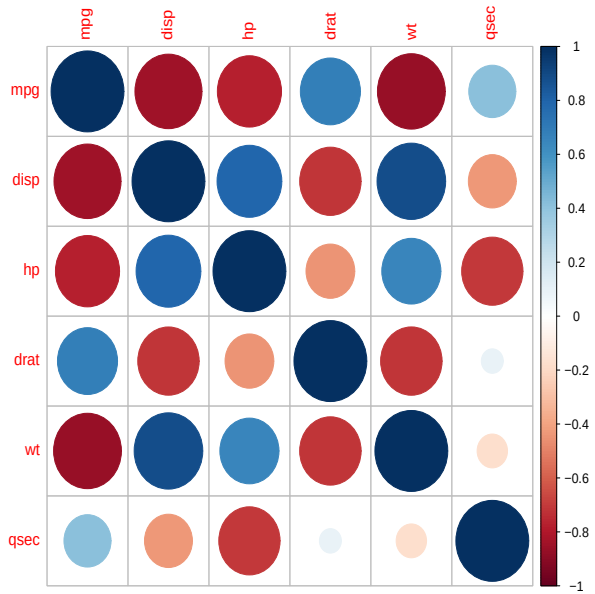
and get

	mpg	disp	hp	drat	wt	qsec
mpg	1.00	-0.85	-0.78	0.68	-0.87	0.42
disp	-0.85	1.00	0.79	-0.71	0.89	-0.43
hp	-0.78	0.79	1.00	-0.45	0.66	-0.71
drat	0.68	-0.71	-0.45	1.00	-0.71	0.09
wt	-0.87	0.89	0.66	-0.71	1.00	-0.17
qsec	0.42	-0.43	-0.71	0.09	-0.17	1.00

- *From the above matrix, the strongest linear predictors of the response* `mpg` *are* `wt` *and* `disp` *since the have the highest correlation with* `mpg`.
- *However, it MUST be noticed that these two are also strongly correlated. Other variables also have strong intercorrelations. These are potential sources of multicollinearity leading to variance inflation.*

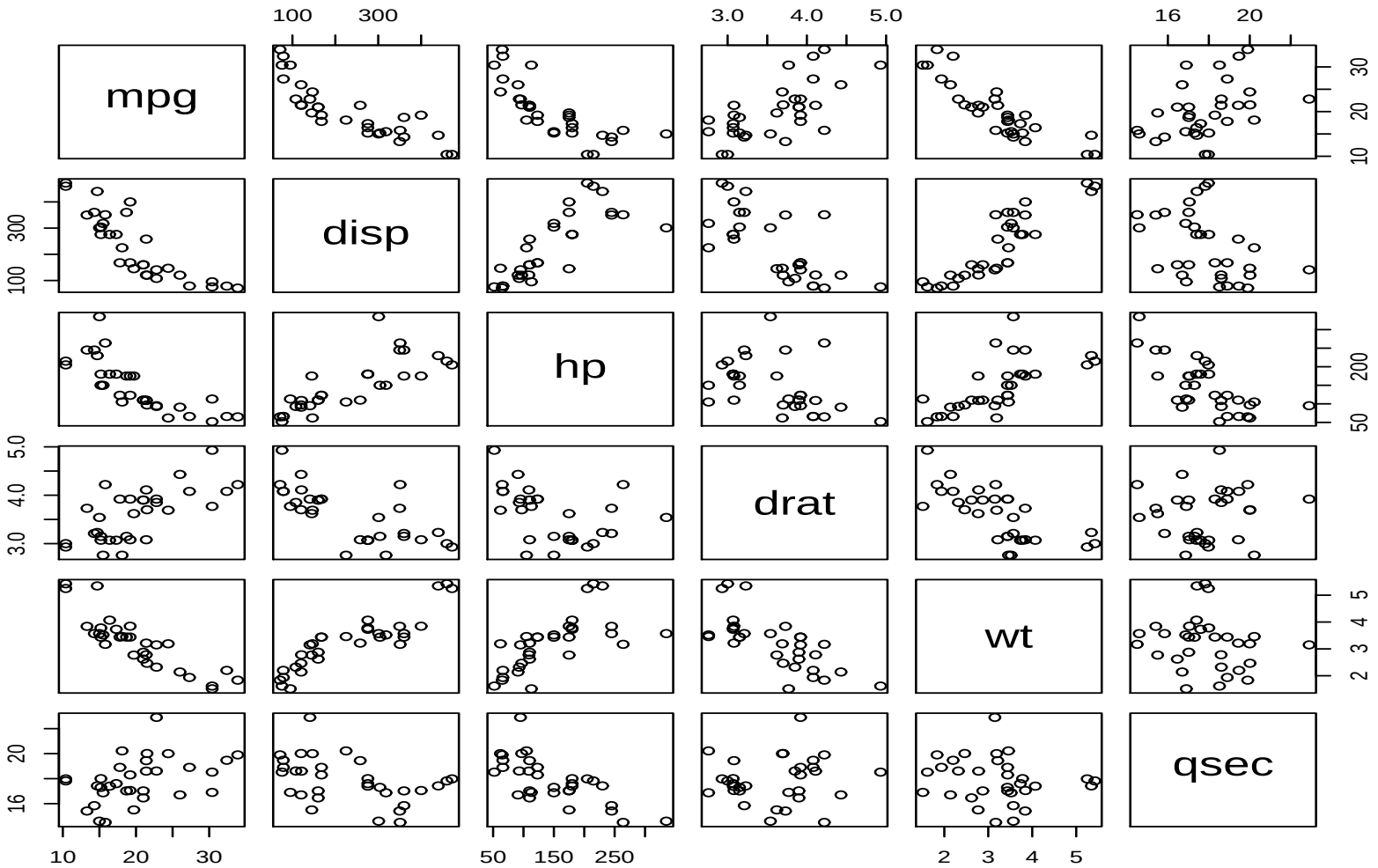
Performing MLR in R

The correlation plot re *correlation matrix*



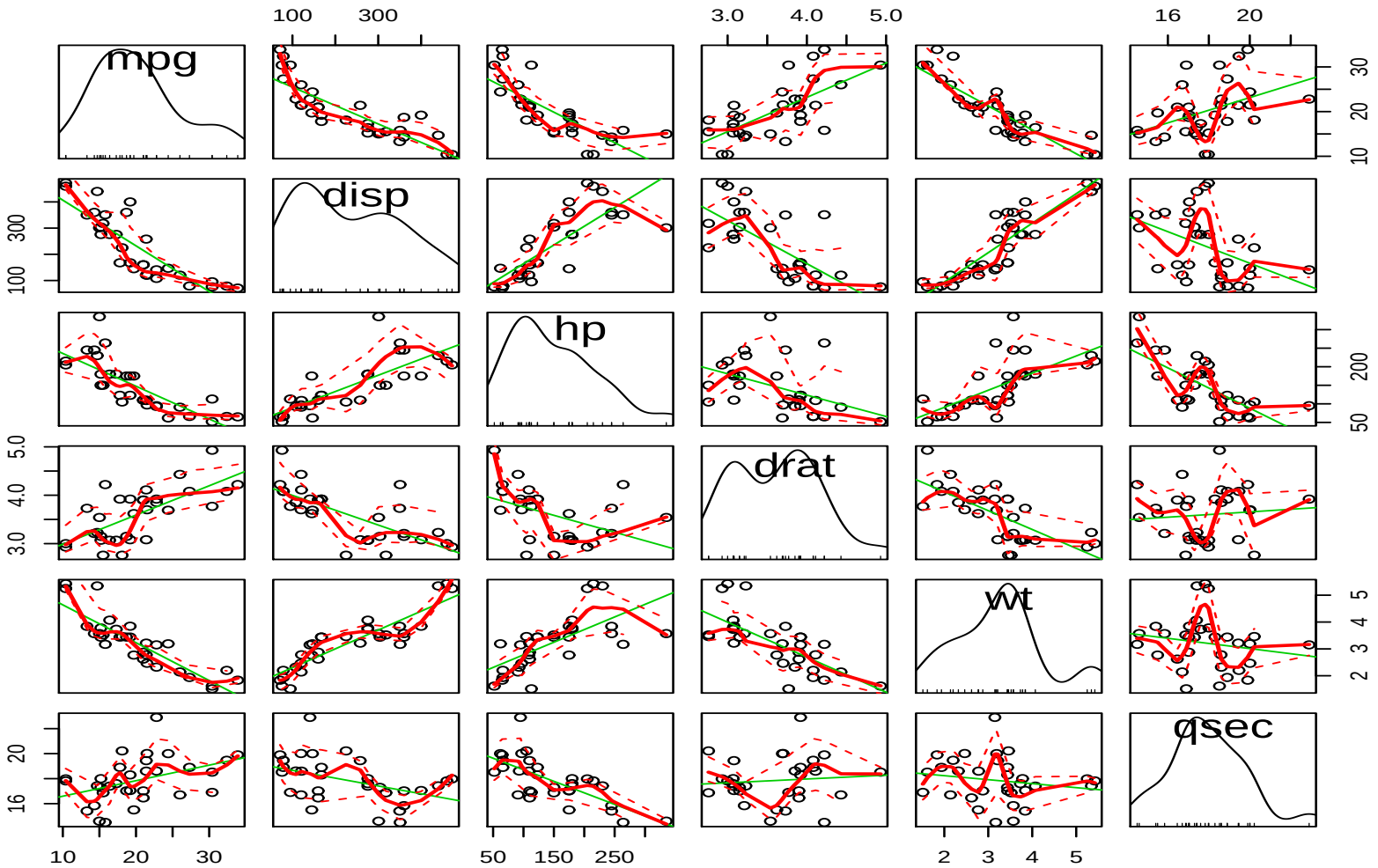
Performing MLR in R

The most basic pairwise scatterplot is crucial to assessing the plausibility of our posited MLR.



Performing MLR in R

This more elaborate scatterplot matrix reveals more clearly the plausibility of aspects of MLR



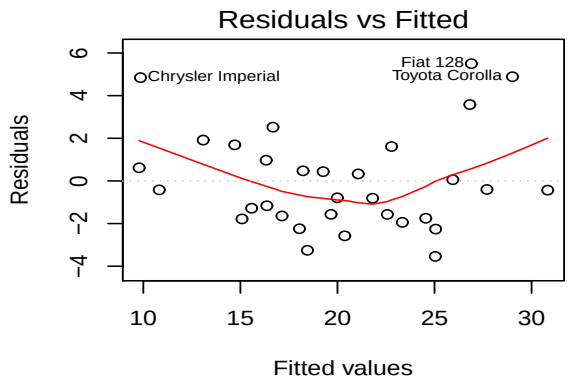
Performing MLR in R

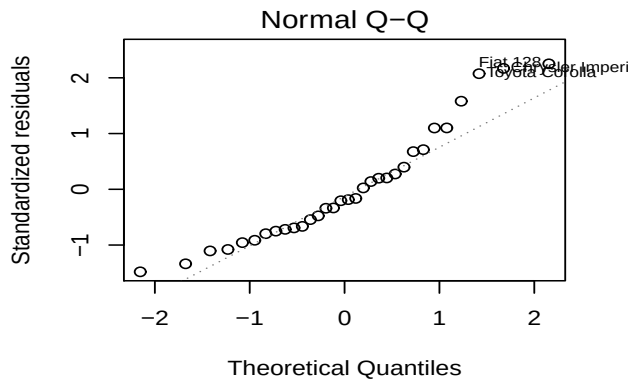
After completing the above steps 1 and 2, we can then fit the model, the summary of which is given by

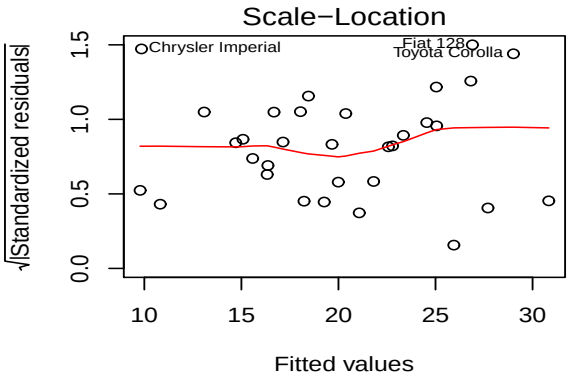
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.53357   10.96423   1.508  0.14362
disp         0.00872    0.01119   0.779  0.44281
hp          -0.02060    0.01528  -1.348  0.18936
drat         2.01578    1.30946   1.539  0.13579
wt          -4.38546    1.24343  -3.527  0.00158 **
qsec         0.64015    0.45934   1.394  0.17523
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.558 on 26 degrees of freedom
Multiple R-squared:  0.8489,    Adjusted R-squared:  0.8199
F-statistic: 29.22 on 5 and 26 DF,  p-value: 6.892e-10
```

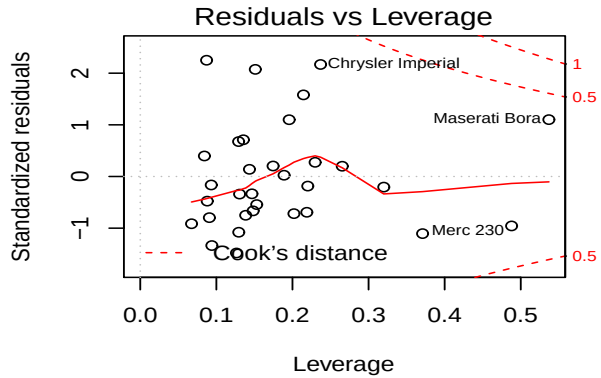
Performing MLR in R on Gas Mileage Data

The following residual analysis plots reveal the fact that a multiple linear regression model doesn't fit. At least there are many aspects that make the MLR model inconclusive









Gas Mileage as a function of Horsepower

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
hp          -0.06823    0.01012  -6.742 1.79e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.863 on 30 degrees of freedom
Multiple R-squared:  0.6024,Adjusted R-squared:  0.5892
F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

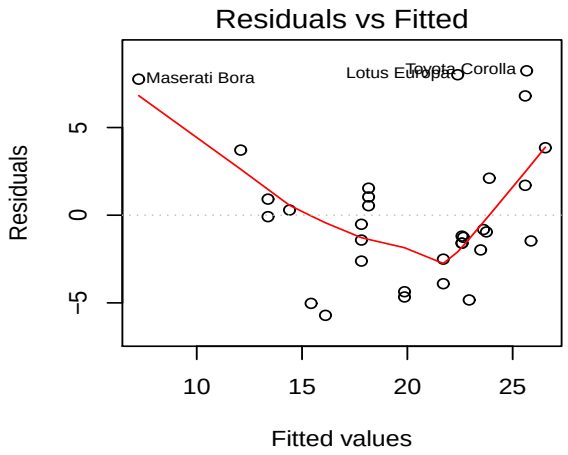
From the above output, the SLR model

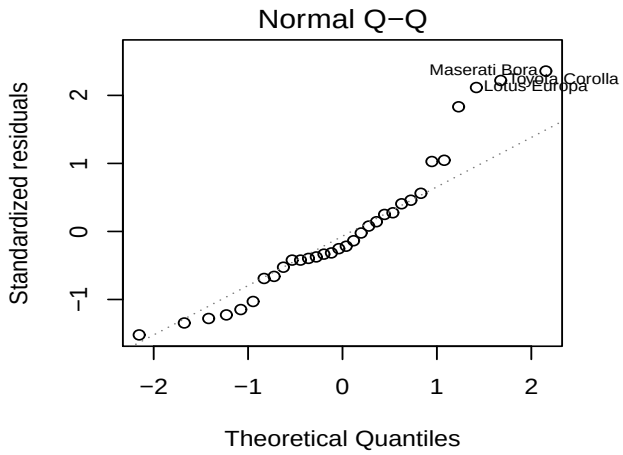
$$\texttt{mpg} = \beta_0 + \beta_1 \texttt{hp} + \epsilon$$

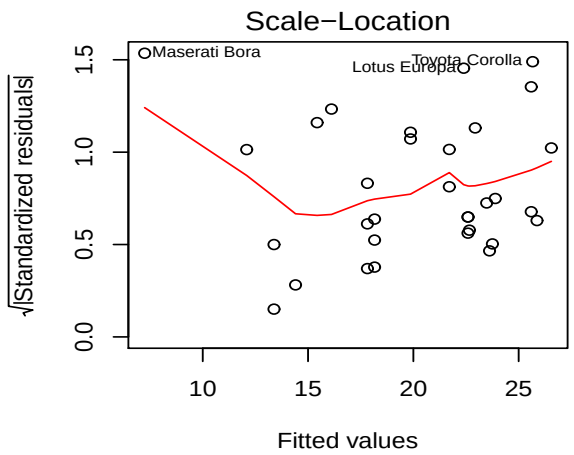
appears significant, judging from the Pvalue. However, does the SLR model provide an adequate fit for the data?

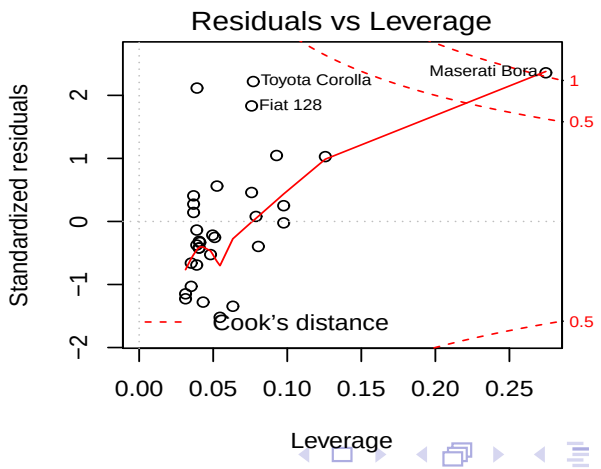
Gas Mileage as a function of Horsepower

The following residual analysis plots further reinforce the fact that a simple linear regression model won't fit well









Gas Mileage as a function of Displacement

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.599855   1.229720  24.070  < 2e-16 ***
disp        -0.041215   0.004712  -8.747 9.38e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.251 on 30 degrees of freedom
Multiple R-squared:  0.7183,    Adjusted R-squared:  0.709
F-statistic: 76.51 on 1 and 30 DF,  p-value: 9.38e-10
```

From the above output, the SLR model

$$\texttt{mpg} = \beta_0 + \beta_1 \texttt{disp} + \epsilon$$

appears significant, judging from the Pvalue. However, does the SLR model provide an adequate fit for the data?

Gas Mileage as a function of Weight

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

From the above output, the SLR model

$$\texttt{mpg} = \beta_0 + \beta_1 \texttt{wt} + \epsilon$$

appears significant, judging from the Pvalue. However, does the SLR model provide an adequate fit for the data?

Gas Mileage as a function of Drat

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    -7.525      5.477  -1.374     0.18
drat            7.678      1.507   5.096 1.78e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.485 on 30 degrees of freedom
Multiple R-squared:  0.464,     Adjusted R-squared:  0.4461
F-statistic: 25.97 on 1 and 30 DF,  p-value: 1.776e-05
```

From the above output, the SLR model

$$\texttt{mpg} = \beta_0 + \beta_1 \texttt{drat} + \epsilon$$

appears significant, judging from the Pvalue. However, does the SLR model provide an adequate fit for the data?

Gas Mileage as a function of Qsec

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.1140    10.0295  -0.510   0.6139
qsec          1.4121     0.5592   2.525   0.0171 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.564 on 30 degrees of freedom
Multiple R-squared:  0.1753,    Adjusted R-squared:  0.1478
F-statistic: 6.377 on 1 and 30 DF,  p-value: 0.01708
```

From the above output, the SLR model

$$\texttt{mpg} = \beta_0 + \beta_1 \texttt{qsec} + \epsilon$$

appears significant, judging from the Pvalue. However, does the SLR model provide an adequate fit for the data?

Foundational Remarks on the MLR fitting results

- *Remark 1: Individual Significance Each of the predictor variables taken alone passed the significance test at $\alpha = 0.05$ as revealed by all the SLR models fitted*
- *Remark 2: Loss of Significance in group Paradoxically, when put together in the MLR model, all the variables fail the significance test, expect for one, namely, wt. To say the least, this is weird.*
- *Remark 3: Multicollinearity This paradoxical phenomenon of a variable both passing and failing is referred to as the paradox of multicollinearity. An immediate source of it in this case is the strong correlation among the predictor variables. In a sense, some variables became redundant in their attempt to explain the response, because another variable strongly correlated with them was already doing the job of explaining the response.*

Foundational Remarks on the MLR fitting results

- *Remark 4: Why can't variables get along* *Well, the answer lies in the mathematics. In the presence of strong correlations among the predictor variables, the crucial variance-covariance matrix of* $\hat{\boldsymbol{\beta}}$ *namely* $\mathbb{V}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$ *becomes severely ill-conditioned, an illness that takes the form of the inflation of the variance of the estimators, results in the loss of significance.*
- *Remark 5: Impact on prediction* *Left unaddressed, multicollinearity leads to unstable models, unreliable inferences, and ultimate erroneous/wrong predictions due to the higher variance than needed of unnecessarily large models.*
- *Remark 6: Remedies* *At the core of this issues, is the non orthogonality of the design matrix* $\mathbf{X}$*. Many remedies have been developed by scientists over the years, all geared towards yielding models such that the variance of* $\hat{\boldsymbol{\beta}}$ *is well-conditioned.*

Foundational Remarks on the MLR fitting results

Among other methods for dealing with multicollinearity, we could mention

- *Principal Component Regression (PCR) that essentially replaced a "bad" $\mathbf{X}$ with an orthogonal and lesser dimensional transformed of itself*
- *Variable Selection which consists of devising strategies and criteria for selecting only those variables that a meaningfully related to the response, rejecting any variable that either not related or correlated to a significant variable already selected*
- *Regularization and Shrinkage methods which consist of foregoing the unbiasedness of $\hat{\boldsymbol{\beta}}^{(\texttt{OLS})}$ and trading-off a bit of bias for a reduction of the variance of $\hat{\boldsymbol{\beta}}$. The famous ridge regression solution is one such solution.*

Simulated Example For Regression Analysis Exploration

Simulated Example: Consider the traditional multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{6}$$

under the assumption that $\epsilon \sim N(0, \sigma^2)$. *We look at a variety of data structures and model selection strategies. Let* $\rho \in [0, 1)$, *then we generate our predictor variables using a multivariate normal distribution with zero mean and the following variance-covariance matrix.*

$$\Sigma = \tau \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{p-1} \\ \rho & 1 & \rho & \cdots & \rho^{p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \cdots & 1 \end{bmatrix}$$

Question: How does the value of ρ *impact/affect the performance of the regression model.*

Exercises and Food for thought

- *Exercise 1: Consider the Prestige dataset from the package car, and use the variable prestige as your response variable. Throughout this study, you should ignore the variable type.*
  1. *Generate the numerical and graphical summaries of each variable including the response, and plot them all on the same panel*
  2. *Compute the correlation matrix, then print it and plot it*
  3. *Generate the scatterplot matrix and comment on the plausibility of a Multiple Linear Regression model for this data*
  4. *Fit all the Simple Linear Regression models and for each SLR, comment on the significance and the goodness of fit*
  5. *Fit the full MLR model and discuss thorough what you see*
  6. *Are there any evidences of potential multicollinearity?*
  7. *Generate the 90% confidence intervals for all the regression coefficients*
  8. *Plot the confidence ellipse of the 1st and 2nd factor,* $\beta_{\texttt{education}}$ *and* $\beta_{\texttt{income}}$
  9. *Generate the 90% confidence intervals for the average mpg of all the cars*
  10. *Generate the 90% prediction intervals for the mpg of all the cars*

Exercises and Food for thought

- *Exercise 2:* Consider the dataset attitude, and perform all the same operations you performed in Exercise 1. Additionally, answer the following questions
  1. Do you have any reason to worry about the scale of measurement of the variables in this case?
  2. If you were to choose a model with two variables, which one would it be and why?
- *Exercise 3:* Find the Boston housing the dataset from R or from the internet, and perform all the same operations you performed in Exercise 1.
- *Exercise 4:* Download the California housing dataset, and comment on the nature of this dataset, especially the way in which it differs from the other datasets explored so far.
- *Exercise 5:* Consider the MLR model

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$. Show that $\mathbb{V}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$.

References

- *Clarke, B, Fokoué, E and Zhang, H (2009).* Principles and Theory for Data Mining and Machine Learning. *Springer Verlag, New York, (ISBN: 978-0-387-98134-5), (2009)*
- *Montgomery, D, Peck, E, and Vining, G (2006).* Introduction to Linear Regression Analysis*(Fourth Edition). Wiley, New York, (ISBN: 0-471-75495-1),(2006)*
- *J. W. Longley (1967).* An appraisal of least-squares programs from the point of view of the user. *Journal of the American Statistical Association, 62, 819-841, (1967)*
- *J. J. Faraway(2002).* Practical Regression and ANOVA using R. *Lecture Notes contributed to the R project, (2002)*