15

LINEAR REGRESSION ANALYSIS

15.1 Linear regression analysis: basics

notations:

- **1** is the vector of all 1.
- J is a square matrix with all 1.

15.1.1 Linear regression models

Linear regression models are arguably the most popular models used to capture the linear relationship between a set of **predictor/regressor variables** , usually denoted by $X_1, X_2, ..., X_k$, and a **response/outcome variable**, denoted by Y. Mathematically, we have

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon,$$

where $\beta_0, ..., \beta_k$ are model coefficients estimated from observations, and ϵ is noise or disturbance.

Standard Linear regression models are usually classified into **simple linear regression model** and **multiple linear regression model**, as follows.

Definition 15.1.1 (simple linear regression model). *The simple linear regression model **assumes** that a random variable Y has a linear dependency on a non-random variable $X \in \mathbb{R}$ given as*

$$Y = \beta_0 + \beta_1 X + \epsilon$$

*where β_0, β_1 are unknown model parameters, and ϵ is a random variable. Given the observed sample pairs $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ as $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and we **further make the following assumptions on** ϵ as*

- $E[\epsilon_i] = 0, \forall i.$
- $Var[\epsilon_j] = \sigma^2, \forall i$; *and σ^2 is unknown.*
- $cov(\epsilon_i, \epsilon_j) = \sigma^2 \delta_{ij}, \forall i, j.$

Definition 15.1.2 (multiple linear regression model). *The multiple linear regression model **assumes** that a random variable Y has a linear dependency on a non-random vector $X = (X_1, X_2, ..., X_{p-1}) \in \mathbb{R}^{p-1}$ given as*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{p-1} X_{p-1} + \epsilon$$

*where $\beta_0, \beta_1, ..., \beta_p$ are unknown model parameters, and ϵ is a random variable. Given the observed sample pairs $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n), x \in \mathbb{R}^{p-1}, y \in \mathbb{R}$ as $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \epsilon_i$ and we **further make the following assumptions on** ϵ as*

- $E[\epsilon_i] = 0, \forall i$
- $cov(\epsilon_i, \epsilon_j) = \sigma^2 \delta_{ij}$ *and* σ^2 *is unknown.*

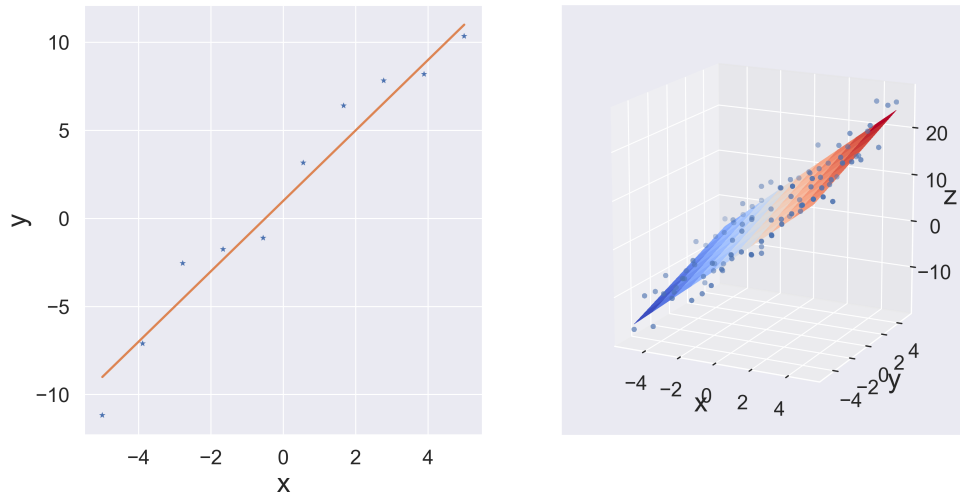


Figure 15.1.1: Demonstration of simple linear regression model $y = \beta_1 x + \beta_0 + \epsilon$ and multiple linear regression model $y = \beta_1 x_1 + \beta_2 x_2 + \beta_0 + \epsilon$. Scatter points are observed data. The solid line in the left and the plane in the right are the mean responses.

The model estimation, interpretation, and improvement in the following sections all rely on model assumptions. Here we reiterate the model assumptions.

Assumption 15.1 (standard assumptions of linear regression model). *[1, p. 17] The standard assumption of a linear regression model consists of*

A1. LINEAR MODEL ASSUMPTION *The random variable Y has a linear dependency on X give by*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{p-1} X_{p-1} + \epsilon.$$

A2. LINEAR INDEPENDENCE OF REGRESSORS *There is no linear dependency exists among regressors $X_1, X_2, ..., X_{p-1}$.*

A3. INDEPEDENCE BETWEEN REGRESSOR AND NOISE *The regressors $(X_1, X_2, ..., X_{p-1})$ are independent of the noise term ϵ.*

A4. HOMOSCEDATICITY *Homoscendaticity refers to that the conditional variance* $Var[\epsilon|X_i], i = 1, ..., p-1$ *is a constant given the observation of the regressors. This can be written by*

$$Var[\epsilon|X_i] = \sigma^2, \forall i.$$

A5. DATA GENERATION OF REGRESSORS *The regressor data* $(x_1, x_2, ...)$ *can be either constants from experimental design or realizations of random variables.*

A6. NORMAL DISTRIBUTION OF NOISE *The noise random samples* $\epsilon_1, \epsilon_2, ...$ *have normal distribution and independent of each other, which can be written by*

$$\epsilon|X_i \sim N(0, \sigma^2 I), \forall i.$$

In the applications of linear regression model, there are usually two types of data generation processes on how we obtain the regressor observations.

- In a physics experiment, the experimentalist will choose different regressor values and observe the output y. In this case, regressor values are certainly not sampled from a distribution.
- In a social study where social scientist usually cannot design experiments like physicist. In this case, we assume regressors are random variables and regressor observations are sampled from a distribution.

Depending on the data generation context, the standard assumption have the following two versions.

With non-random x	With random x
A1: $y = \beta_1 + x^T\beta_2 + e$, with x fixed	A1: $y = \beta_1 + x^T\beta_2 + e$, with x, e random
A2: $E[e] = 0$	A2: $E[e] = 0$
A3: $Var[e] = \sigma^2$	A3: $Var[e] = \sigma^2$
A4: $Cov(e_i, e_j) = 0$	A4: $Cov(e_i, e_j) = 0$
A5: x	A5: x
A6: e is normal	A6: e is normal

In the theoretical treatment of linear regression models, we assume predictor variables are taking continuous values. In practice, regressors variables can take discrete values. Suppose we have regressor characterizing whether tomorrow is rainy or not. Then we can design a regressor with the following rule

$$x_i = \begin{cases} 1, rainy \\ 0, not\ rainy \end{cases}$$

And our linear regression theory will accommodate such binary variable. The creation of proxy binary variable for discrete variables taking $K, K \geq 3$ discrete values is via similar procedure [subsubsection 22.5.1.3].

15.1.2 Ordinary least square (OLS): fundamentals

15.1.2.1 *Review on orthogonal projections*

We will primarily take a geometric approach to linear regression analysis. Before we get into OLS results, we first review the properties of an orthogonal projector in the form of $X(X^TX)^{-1}X$. Its properties can be extensively and repeatedly in the following development.

Lemma 15.1.1 (essential properties of orthogonal projection). *Let X be a matrix of size $n \times p$. Define*

$$H = X(X^TX)^{-1}X^T,$$

we have

- *$(X^TX)^{-1}$ exists if X has full column rank.*
- *$rank((X^TX)) = p$, $rank(H) = p$ and $rank(I - H) = n - p$.*
- *H is symmetric and idempotent; in other words, H is an orthogonal projector onto the subspace spanned by columns of X.*
- *$I - H$ is symmetric and idempotent; in other words, $I - H$ is an orthogonal projector onto the orthogonal complementary subspace spanned by columns of X.*

Proof. (1) Use the fact that $\mathcal{N}(X^TX) = \mathcal{N}(X)$. If X is invertible then X^TX is invertible. (2) X^TX is invertible, therefore has full rank of p. Use Lemma 4.4.1.

$$rank(X(X^TX)^{-1}) = rank((X^TX)^{-1}) = p, rank(X(X^TX)^{-1}X^T) = rank((X^TX)^{-1}) = p.$$

(3) Direct verification. Theorem 4.5.6. (4) $(I - H)^T = I - H$, and $(I - H)(I - H) = (I - 2H + H^2) = (I - 2H + H) = I - H$. □

15.1.2.2 *OLS results*

In linear regression analysis, estimation of model coefficients β via least square minimization of predicted error gives the following fundamental result.

Theorem 15.1.1 (fundamental least square solution: general case). *The multiple linear regression with n samples can be written as*

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \dots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n(p-1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

with matrix form

$$Y = X\beta + \epsilon.$$

Assume the standard assumptions [Assumption 15.1] hold. The ***unique minimizer*** *to the problem*

$$\min_{\beta} (Y - X\beta)^T (Y - X\beta)$$

is given by

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

in particular,

$$\hat{\beta}_0 = \overline{y} - \sum_{i=1}^{p-1} \hat{\beta}_i \overline{x}_i$$

$$\hat{y}_i - \overline{y} = \sum_{i=1}^{p-1} \hat{\beta}_j (x_{ij} - \overline{x}_j)$$

Moreover, we have

- $E[\hat{\beta}] = \beta$.
- *If* $Cov[Y] = \sigma^2 I$, *then* $Cov[\hat{\beta}] = \sigma^2 (X^T X)^{-1}$.($\hat{\beta}$ *is not necessarily normal).*
- *To get each individual coefficient, we have*

$$\hat{\beta}_i = \frac{X_i^T (I - H_{-i}) Y}{X_i^T (I - H_{-i}) X_i},$$

where $H_{-i} = X_{-i}(X_{-i}^T X_{-i})^{-1} X_{-i}^T$, X_{-i} *is the matrix without column i.*

Proof. $\hat{\beta}$ can be obtained via direct optimization or using normal equation theorem from Theorem 5.4.4.
(1) (unbiased)

$$E[\hat{\beta}] = (X^T X)^{-1} X^T E[Y] = (X^T X)^{-1} X^T X\beta = \beta.$$

(2) (variance)

$$Cov[\beta] = (X^TX)^{-1}X^TCov[Y]((X^TX)^{-1}X^T)^T = \sigma^2(X^TX)^{-1}$$

. (3) See Theorem 5.4.4. We can interpret as first projecting Y into the null space of $(I - H_{-i})X_{-i}$ and project onto X_i. □

Remark 15.1.1 (geometric interpretation).

- We can use projection theorem in Hilbert space to interpret the result. We can treat the the observation Y as a vector in $\mathbb{R}^N$, and the observations of $1, X_1, X_2, .., X_k$ form a linear subspace. And we are trying to find the minimizing vector β as the projection of Y onto the subspace.
- Indeed, the prediction $\hat{Y} = X(X^TX)^{-1}X^TY$ exactly present the projection of Y onto the column subspace of X.
- The Gramm matrix X^TX is always semi-positive definite but might be ill-conditioned due to the linear dependence of columns(which suggests some features are unnecessary). The linear dependence of columns can be checked using sample correlations. We can also use SVD to reduce dimensionality.

Remark 15.1.2 (change of coefficient value after adding or omitting regressors). Consider a original linear regression model given by

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1}X_{k-1}.$$

Suppose now we add a new regressor X_k. Then the coefficient $\hat{\beta}_k$ has the following change after adding the new regressor

- If X_k is demeaned and uncorrelated with $X_1, X_2, ..., X_{k-1}$, then the coefficient $\hat{\beta}_i$ will not change.
- If X_k has finite mean and uncorrelated with $X_1, X_2, ..., X_{k-1}$, then the coefficient $\hat{\beta}_i, i = 1, 2, ..., k-1$ will not change, coefficient $\hat{\beta}_0$ will decrease if X_k is positively correlated with Y and has positive mean (i.e., positively correlated with unit vector $\mathbf{1}$), and coefficient $\hat{\beta}_0$ will increase if X_k is positively correlated with Y and has negative mean (i.e., negatively correlated with unit vector $\mathbf{1}$). A simple intuition is

$$\beta_0 = E[Y] - \beta_1 E[X_1] - \cdots - \beta_k E[X_k].$$

- More generally, if X_k is arbitrary, then the coefficient $\hat{\beta}_i$ will decrease if $Corr(X_k, Y)$ and $Corr(X_k, X_i)$ have the same sign (competing effect) and the coefficient $\hat{\beta}_i$ will increase if $Corr(X_k, Y)$ and $Corr(X_k, X_i)$ have the opposite (compensating effect). For example, suppose $Corr(X_k, Y) > 0$, then $\hat{\beta}_k > 0$. Because $Corr(X_k, X_i) < 0$, $\hat{\beta}_i$ needs to increase to compensate the offsetting effects due to X_k.

Corollary 15.1.1.1 (special cases, least square solution for simple regression). *Define*

$$S_{XX} = \sum_i (x_i - \overline{x})^2, S_{XY} = \sum_i (x_i - \overline{x})(y_i - \overline{y}, S_{YY} = \sum_i (y_i - \overline{y})^2,$$

and

$$S_X = \sqrt{S_{XX}}, S_Y = \sqrt{S_{YY}}.$$

It follows that

- *For the zero order model $y = \beta_0 + \epsilon$,*

$$\hat{\beta}_0 = \frac{\langle y, \mathbf{1} \rangle}{\langle \mathbf{1}, \mathbf{1} \rangle} = \frac{1}{n} \sum_i y_i$$

- *For the first order model $y = \beta_0 + \beta_1 x + \epsilon$,*

$$\hat{\beta} = [X^T X]^{-1} X^T Y$$

we have

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

and

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} = \frac{S_{XY}}{S_{XX}} = \hat{\rho} \frac{S_Y}{S_X}$$

where $\hat{\rho} = S_{XY}/S_X S_Y$ is the sample correlation coefficient between X and Y. In summary,

$$\hat{y}_i = \overline{y} + \hat{\beta}_1 (x_i - \overline{x}).$$

- *In the first order model, we have conditional mean and unconditional variance of y given by*

$$E[y|x] = \beta_0 + \beta_1 x, Var[y] = \sigma^2, Var[\overline{y}] = \frac{\sigma^2}{n}.$$

- *In the first order model, we have unbiased coefficient estimator*

$$E[\hat{\beta}_1] = \beta_1, E[\beta_0] = \beta_0.$$

- *In the first order model, we have*

$$Var[\hat{\beta}_1] = \frac{\sigma^2}{S_{XX}}, Var[\hat{\beta}_0] = \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}), Cov(\hat{\beta}_1, \hat{\beta}_0) = -\frac{\sigma^2\overline{x}}{S_{XX}}$$

where

$$S_{XX} = \sum_i (x_i - \overline{x})^2 = \sum_i (x_i - \overline{x})x_i.$$

Proof. (2)

$$Var[y] = E[(y - E[y])^2] = E[(y - E[y|x])^2] = E[\epsilon^2] = \sigma^2.$$

(3)Note that

$$\begin{aligned}\sum_i (x_i - \overline{x})^2 &= x^T(I - \frac{1}{n}J)(I - \frac{1}{n}J)x \\ &= x^T(I - \frac{1}{n}J)x \\ &= x^T(x - \overline{x}\mathbf{1})\end{aligned}$$

and

$$\begin{aligned}\sum_i (x_i - \overline{x})(y_i - \overline{y}) &= x^T(I - \frac{1}{n}J)(I - \frac{1}{n}J)y \\ &= x^T(I - \frac{1}{n}J)y \\ &= x^T(y - \overline{y}\mathbf{1}) \\ &= y^T(x - \overline{x}\mathbf{1})\end{aligned}$$

(4)

$$\begin{aligned}E[\hat{\beta}_1] &= E[\frac{\sum_i (x_i - \overline{x})y_i}{\sum_i (x_i - \overline{x})(x_i)}] \\ &= E[\frac{\sum_i (x_i - \overline{x})(\beta_0 + \beta_1 x_i + \epsilon_i)}{\sum_i (x_i - \overline{x})(x_i)}] \\ &= \frac{\sum_i (x_i - \overline{x})(\beta_0 + \beta_1 x_i)}{\sum_i (x_i - \overline{x})(x_i)} \\ &= \frac{\sum_i (x_i - \overline{x})(\beta_0)}{\sum_i (x_i - \overline{x})(x_i)} + \frac{\sum_i (x_i - \overline{x})(\beta_1 x_i)}{\sum_i (x_i - \overline{x})(x_i)} \\ &= 0 + \beta_1 = \beta_1.\end{aligned}$$

and

$$\begin{aligned} E[\hat{\beta}_0] &= E[\overline{y} - \hat{\beta}_1 \overline{x}] \\ &= \frac{1}{n} \sum_i E[y_i - \hat{\beta}_1 x_i] \\ &= \frac{1}{n} \sum_i E[\beta_0 + \beta_1 x_i - \hat{\beta}_1 x_i] \\ &= \frac{1}{n} \sum_i \beta_0 \\ &= \beta_0. \end{aligned}$$

(5)

$$\begin{aligned} Var[\hat{\beta}_1] &= Var[\frac{\sum_i (x_i - \overline{x}) y_i}{\sum_i (x_i - \overline{x})^2}] \\ &= Var[\frac{\sum_i (x_i - \overline{x})(\beta_0 + \beta_1 x_i + \epsilon_i)}{\sum_i (x_i - \overline{x})^2}] \\ &= Var[\frac{\sum_i (x_i - \overline{x})(\epsilon_i)}{\sum_i (x_i - \overline{x})(x_i)}] \\ &= \sigma^2 \sum_i \frac{\sum_i (x_i - \overline{x})^2}{(\sum_i (x_i - \overline{x})^2)^2} \\ &= \sigma^2 \frac{1}{(\sum_i (x_i - \overline{x})^2)} \end{aligned}$$

and

$$\begin{aligned} Var[\hat{\beta}_0] &= Var[\overline{y} - \hat{\beta}_1 \overline{x}] \\ &= Var[\overline{y}] + \overline{x}^2 Var[\hat{\beta}_1] - 2\overline{x} Cov(\overline{y}, \hat{\beta}_1) \\ &= \frac{1}{n}\sigma^2 + \overline{x}^2 \sigma^2 / S_{XX} + 0 \\ &= \sigma^2 (\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}) \end{aligned}$$

where $Cov(\overline{y}, \hat{\beta}_1) = 0$ since

$$\begin{aligned} Cov(\overline{y}, \hat{\beta}_1) &= E[\frac{1}{n} \sum_i \epsilon_i (\sum_i c_i \epsilon_i - \beta_1)] \\ &= \frac{1}{n} E[\sum_i c_i \epsilon_i^2] \\ &= \frac{1}{n} \sum_i c_i \sigma^2 \\ &= 0 \end{aligned}$$

where

$$c_i = \frac{(x_i - \overline{x})}{\sum_i (x_i - \overline{x})(x_i)}, \sum_i c_i = 0.$$

To calculate $Cov(\hat{\beta}_1, \hat{\beta}_0)$, we use the fact that

$$\begin{aligned} \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \\ \hat{\beta}_0 + \hat{\beta}_1 \overline{x} &= \overline{y} \\ Var[\hat{\beta}_0 + \hat{\beta}_1 \overline{x}] &= Var[\overline{y}] = \frac{\sigma^2}{n} \\ Var[\hat{\beta}_0] + Var[\hat{\beta}_1]\overline{x}^2 + 2\overline{x} Cov(\hat{\beta}_1, \hat{\beta}_0) &= \frac{\sigma^2}{n} \end{aligned}$$

then we can get

$$Cov(\hat{\beta}_1, \hat{\beta}_0) = -\frac{\sigma^2 \overline{x}}{S_{XX}}.$$

□

15.1.2.3 *OLS results with demeaned data*

In many applications, data will be standardized, including removing mean, before running least square analysis. It is curious to know whether removal of data mean will affect the OLS estimate. The following results show that the coefficient estimate remains the same.

Theorem 15.1.2 (least square solution: demean case). *Consider a multiple linear regression problem where y and x are demeaned; that is, $\sum_i y_i = 0, \sum_i x_{i,j} = 0, j = 1, 2, ..., p$. The multiple linear regression with n samples can be written as*

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_0 \\ \vdots \\ \beta_0 \end{bmatrix} + \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1(p)} \\ x_{21} & x_{22} & \dots & x_{2(p)} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{n(p)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

with matrix form

$$Y = \beta_0 \mathbf{1} + X\beta + \epsilon$$

*The **unique minimizer** to the problem*

$$\min_{\beta} (Y - \beta_0 \mathbf{1} - X\beta)^T (Y - \beta_0 \mathbf{1} - X\beta)$$

is given as

$$\hat{\beta}_0 = \overline{y} - \sum_{i=1}^{p-1} \hat{\beta}_i \overline{x}_i,$$

and

$$\hat{\beta} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y},$$

where $\tilde{X}_{ij} = X_{ij} - \overline{x}_j, \tilde{Y} = Y - \overline{y}$.

Eventually, we can write

$$\hat{y}_i - \overline{y} = \sum_{j=1}^{p-1} \hat{\beta}_j (x_{ij} - \overline{x}_j)$$

Moreover, we have

$E[\beta_0] = \beta_0, E[\hat{\beta}] = \beta$.*ditemize*

Proof. (1)We can use the projection theorem Theorem 5.4.4 or use the following optimization method.

$$\min f = (Y - \beta_0 \mathbf{1} - X\beta)^T (Y - \beta_0 \mathbf{1} - X\beta)$$

over β_0, β_1, we have

$$\begin{aligned} f(\beta_0, \beta_1) &= Y^T Y + n^2 \beta_0^2 + (\beta^T X^T X \beta) + 2n\beta_0 \mathbf{1}^T X\beta - 2n\beta_0 \mathbf{1}^T Y - 2Y^T X\beta \\ &= Y^T Y + n^2 \beta_0^2 + (\beta^T X^T X \beta) + 2n\beta_0 \mathbf{1}^T X\beta - 2n\beta_0 \sum_{i=1}^{n} y_i - 2Y^T X\beta \end{aligned}$$

The first order condition on β_0 gives that

$$\beta_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - \frac{1}{n} \mathbf{1}^T X\beta;$$

Plug in β_0(note that $(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)X = \tilde{X}$), we can transform the minimization problem to

$$f(\beta_0, \beta_1) = (\tilde{Y} - \tilde{X}\beta)^T (\tilde{Y} - \tilde{X}\beta).$$

Then we can use the results in the general case [Theorem 15.1.1] to obtain the estimator for β. (2) (unbiasedness)

$$\begin{aligned}
E[\hat{\beta}_0] &= E[\overline{y} - \sum_{i=1}^{p-1} \hat{\beta}_i \overline{x}_i] \\
&= E[\overline{y}] - \sum_{i=1}^{p-1} E[\hat{\beta}_i \overline{x}_i] \\
&= E[\frac{1}{n}\mathbf{1}^T Y] - \sum_{i=1}^{p-1} E[\hat{\beta}_i]\overline{x}_i \\
&= \frac{1}{n}\mathbf{1}^T E[Y] - \sum_{i=1}^{p-1} \frac{1}{n}\mathbf{1} X_i \beta_i \\
&= \frac{1}{n}\mathbf{1}^T (Y - X\beta) \\
&= \frac{1}{n}\mathbf{1}^T (\beta_0 \mathbf{1} + X\beta + \epsilon - X\beta) \\
&= \frac{1}{n}\mathbf{1}^T \beta_0 \mathbf{1} \\
&= \beta_0.
\end{aligned}$$

where we use the fact that $E[\hat{\beta}] = \beta$ proved later.

$$E[\hat{\beta}] = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T E[Y] = (X^T X)^{-1} \tilde{X}^T \tilde{X} \beta = \beta.$$

□

Remark 15.1.3 (another way to see the coefficients in the demean case). From Theorem 15.1.1, we can see that if we want to get the coefficients $\beta = (\beta_1, \beta_2, ..., \beta_{p-1})$, we can use

$$\beta = (X^T (I - H_0)(I - H_0) X)^{-1} X^T (I - H_0)(I - H_0) Y,$$

where $H_0 = \frac{1}{n}\mathbf{1}\mathbf{1}^T$. Note that $(I - H_0)Y$ will generate a demeaned Y and $(I - H_0)X$ will generate a column-wise demeaned X. This is because the $I - H_0$ operator will is a demean operator.

Remark 15.1.4 (interpretation).

- We can interpret $x_i - \overline{x}$ and $y_i - \overline{y}$ as the part remove the projection (that is, $\overline{x}\mathbf{1}, \overline{y}\mathbf{1}$)onto the constant value subspace of $\mathbf{1}$. To see this, we have

$$\sum_{i=1}^{n} (x_i - \overline{x})\overline{x} = n\overline{x}^2 - n\overline{x}^2 = 0,$$

and

$$\sum_{i=1}^{n}(y_i - \overline{y})\overline{y} = n\overline{y}^2 - n\overline{y}^2 = 0.$$

Therefore, the formula

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

is consistent with the formula

$$\hat{\beta}_i = \frac{X_i^T (I - H_{-i})Y}{X_i^T (I - H_{-i})X_i},$$

where $H_{-i} = X_{-i}(X_{-i}^T X_{-i})^{-1} X_{-i}^T$, X_{-i} is the matrix without column i.

- The coefficient $\hat{\beta}_j$ represents the additional contribution from X_j after accounting for the contribution from $\mathbf{1}, X_1, ..., X_{j-1}, X_{j+1}, ..., X_p$.

15.1.2.4 *Gauss-Markov theorem*

OLS estimate is just one way of estimating model parameters. How does OLS estimate compare with other estimators. The Gauss–Markov theorem we are covering now states that under standard assumption of linear regression, OLS estimate actually gives the best linear unbiased estimator (BLUE). Here *best* means the estimator gives the lowest variance of the estimate, as compared to other unbiased, linear estimators.

If we further assume the error follows normal distribution, the OLS is indeed the best unbiased estimator.

Theorem 15.1.3 (Gauss-Markov theorem, best linear unbiased estimator (BLUE)). *Given the statistical model*

$$Y = X\beta + \epsilon, E[\epsilon] = 0, Cov(\epsilon) = \sigma^2 I$$

with β being the model parameter, y being the observations, the ***uniformly minimum variance estimators among all linear unbiased estimators*** *is given by*

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

As a summary, we have

- $E[\hat{\beta}] = \beta$.
- $Cov[\hat{\beta}] = \sigma^2 (X^T X)^{-1}$.

- *Furthermore, if ϵ is Gaussian noise, i.e., $\epsilon \sim MN(0, \sigma^2 I)$, and $Y, X_1, X_2, ..., X_n$ are multivariate Gaussian, then $\hat{\beta}$ is the uniformly minimum variance estimator among all estimators.*

Proof. (1) (unbiased) $E[\beta] = (X^T X)^{-1} X^T E[Y] = (X^T X)^{-1} X^T X\beta = \beta$. (2) Let $\theta' = AY$ be any other unbiased linear estimator, and assume $\theta' = (X^T X)^{-1} X^T + D$ for some matrix D. The unbiasedness requires that

$$E\theta' = \theta(I + DX) = \theta \Rightarrow DX = 0.$$

The variance of the estimator is given as

$$\begin{aligned} E[(\theta' - \theta)(\theta' - \theta)^T] &= E[(D + (X^T X)^{-1} X^T)\epsilon\epsilon^T (D + (X^T X)^{-1} X^T)^T] \\ &= DE[\epsilon\epsilon^T]D^T + (X^T X)^{-1}\sigma^2 I \\ &= \sigma^2(DD^T + (X^T X)^{-1}) = \sigma^2 DD^T + Var(\theta) \geq Var(\theta). \end{aligned}$$

Here the $\geq$ sign is in the semi-positive matrix sense. (3) Note that the joint pdf of $y = (y_1, ..., y_n)$ can be written by

$$f(y; \beta, \sigma^2) = \prod_{i=1}^{N} N(y_i; x_i, \beta, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp(-\frac{1}{2\sigma^2}(y_i - x_i^T\beta)^T(y_i - x_i^T\beta)).$$

The Fisher information associated with f is given by

$$\frac{\partial \log f}{\partial \beta} = \frac{1}{\sigma^2}(Y - X\beta)^T X.$$

$$I(\beta) = -E[\frac{\partial^2 \log f}{\partial\beta\partial\beta^T}] = \frac{1}{\sigma^2} X^T X.$$

It is clear that $Cov[\hat{\beta}] = [I(\beta)]^{-1}$; that is, the variance of $\hat{\beta}$ reaches the Cramer-Rao lower bound [Theorem 13.2.5]. Therefore, $\hat{\beta}$ has the uniformly minimum variance. □

Remark 15.1.5 (the distribution of noise and its consequence).

- The noise ϵ does not need to be Gaussian, but required to have zero mean and σ^2 variance. In this case, we get the best linear estimator among all the linear estimators.
- If the noise follows Gaussian distribution and random variable $Y, X_1, X_2, ..., X_n$ are multivariate Gaussian, then we get the best estimator among all the estimators.

15.1.2.5 *Variance decomposition*

Theorem 15.1.4 (variance decomposition and property in linear regression with normality assumption). *In the linear regression(with p coefficients), let Y be the vector of the observations, and let H be the orthogonal projector of rank p and $\epsilon \sim MN(0, \sigma^2 I)$. Then we have*

- *$I - \frac{1}{n}J$, $I - H$, and $H - \frac{1}{n}J$ are orthogonal projectors with rank $n-1$, $n-p$, and $p-1$, respectively. And they are mutually orthogonal to each other.*
- $$\begin{aligned} SST &\triangleq \sum_{i=1}^{n}(y_i - \overline{y})^2 \\ &= Y^T(I - \frac{1}{n}J)Y \\ &\sim \sigma^2\chi^2(n-1) \end{aligned}$$
- $$\begin{aligned} SSE &\triangleq \sum_{i=1}(y_i - \hat{y}_i)^2 \\ &= Y^T(I - H)Y \\ &\sim \sigma^2\chi^2(n-p) \end{aligned}$$
- $$\begin{aligned} SSR &\triangleq \sum_{i=1}(\hat{y}_i - \overline{y})^2 \\ &= Y^T(H - \frac{1}{n}J)Y \\ &\sim \sigma^2\chi^2(p-1) \end{aligned}$$
- ***(independence)*** *SSE and SSR are independent.*
- ***(partition identity)***
$$SST = SSE + SSR$$
- *$\hat{\beta} = (X^TX)^{-1}X^TY$ and $(I-H)Y$ are mutually independent from each other; moreover, $\hat{\beta}$ and $SSE = Y^T(I-H)Y$ are independent from each other.*

Proof. (1) From linear algebra theory [], a projector P is orthogonal projector if $P^2 = P$ and $P^T = P$. It is easy to see that these three projectors are symmetric. It is not hard to show that

$$(I - \frac{1}{n}J)^T(I - \frac{1}{n}J) = (I - \frac{1}{n}J),$$
$$(I - H)^T(I - H) = (I - H),$$

and

$$\begin{aligned}(H-\frac{1}{n}J)^T(H-\frac{1}{n}J) &= H^2 - H\frac{2}{n}J + \frac{1}{n^2}J^TJ \\ &= H - \frac{2}{n}J + \frac{1}{n}J \\ &= H - \frac{1}{n}J\end{aligned}$$

where we used the fact that $H\frac{2}{n}J = \frac{2}{n}J$ because the subspace associated with H contains the subspace associated with $\frac{2}{n}J$.

In addition, we can similarly show

$$(I-\frac{1}{n}J)^T(H-\frac{1}{n}J) = 0, (I-H)^T(H-\frac{1}{n}J) = 0, (I-H)^T(I-\frac{1}{n}J) = 0.$$

(2) Since $\epsilon \sim MN(0, \sigma^2 I)$, y_i is a normal random variable. Therefore $SST = Y^T(I - \frac{1}{n}J)Y$ is just the sample variance for normal random samples. It has been showed that sample variance follows $SST \sim \sigma^2\chi^2(n-1)$[Corollary 12.4.3.1]. (3)(4)(5) Note that H has rank p [Lemma 15.1.1]. And $Y^T(I-\frac{1}{n})Y = Y^T(I-H)Y + Y^T(H-\frac{1}{n})Y$. Then via Theorem 12.4.3 we know that

$$SST = SSE + SSR, SSE \sim \sigma^2\chi^2(n-p), SSR \sim \sigma^2\chi^2(p-1).$$

(6) From Lemma 12.4.3, we have

$$\begin{aligned}(I-H)(H-\frac{1}{n}J) &= H - \frac{1}{n}J - H^2 - \frac{1}{n}HJ \\ &= H - \frac{1}{n}J - H - \frac{1}{n}J \\ &= 0\end{aligned}$$

where we used the fact that $H\frac{1}{n}J = \frac{1}{n}J$ because the subspace associated with H contains the subspace associated with $\frac{1}{n}J$. (7) Note that under the assumption of $\epsilon \sim MN(0, I)$, $\hat{\beta}$ and $(I-H)Y$ are multivariate normal random vectors. To show independence, we have

$$\begin{aligned}&Cov((X^TX)^{-1}X^TY, (I - X(X^TX)^{-1}X^T)Y) \\ =&(X^TX)^{-1}X^TCov(Y,Y)(I - X(X^TX)^{-1}X^T) \\ =&(X^TX)^{-1}X^T - (X^TX)^{-1}X^T = 0.\end{aligned}$$

Since $\hat{\beta}$ and $(I-H)Y$ are independent, we can directly see $\hat{\beta}$ and $Y^T(I-H)(I-H)Y$ are independent. □

Theorem 15.1.5 (residuals and estimation of variance for linear regressions). *[2, p. 114] Let $SSE = \sum_{i=1}(y_i - \hat{y}_i)^2$ be the residual of a p-order (simple linear regression corresponds to $p = 2$) linear regression with n samples.*

- *One unbiased estimator of σ^2 is given by*

$$\hat{\sigma}^2 = \frac{SSE}{n-p}$$

- *If $\epsilon \sim MN(0, \sigma^2 I)$, then*

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-p).$$

from which, we can estimate σ via

$$\hat{\sigma}^2 = \frac{SSE}{n-p},$$

where we have

$$E[\hat{\sigma}^2] = E[\frac{SSE}{n-p}] = \sigma^2,$$

and

$$Var[\hat{\sigma}^2] = \frac{\sigma^4}{n-p}.$$

Proof. (1) Note that

$$SSE = Y^T(I-H)Y = \epsilon^T(I-H)\epsilon,$$

since $(I-H)X = 0$. Then

$$E[SSE] = E[\epsilon^T(I-H)\epsilon] = E[Tr((I-H)\epsilon\epsilon^T)] = Tr((I-H)E[\epsilon\epsilon^T]) = \sigma^2(n-p),$$

where we use the fact that the orthogonal projector $(I-H)$ has rank $n-p$ [Lemma 15.1.1]. (2) use Theorem 15.1.4; then use the properties of χ^2 [Lemma 12.1.32]. □

15.1.2.6 *Residual and variance estimation*

After fitting the model parameter $\hat{\beta}$, we can further perform residual analysis and estimate variance of the noise.

Definition 15.1.3 (residual of a linear regression). *Let $\hat{\beta} = (X^T X)^{-1} X^T Y$ be the OLS estimate such that $\hat{Y} = X(X^T X)^{-1} X^T Y = HY$. The **residual** of linear regression is defined as*

$$SSE = \sum_{i=1} (y_i - \hat{y}_i)^2 = Y^T (I - H) Y$$

where $H = X(X^T X)^{-1} X^T$ is the orthogonal projector associated with the linear regression.[a]

[a] The residual is given in matrix form as

$$SSE = (Y - HY)^T (Y - HY) = Y^T Y - Y H^T H Y - 2 Y^T H Y = Y^T I Y - Y^T H Y$$

where the orthogonality property of H,$H^T = H, H^2 = H$ is used.

Lemma 15.1.2 (basic property of residual vector). *Let $e = (I - H)Y$ be the residual vector. We have*

- $$X^T e = 0;$$

  That is, residual vector is orthogonal to the subspace of observations.
- $$\hat{Y}^T e = 0;$$

  That is, residual vector is orthogonal to the projection of Y.

Proof. (1)

$$X^T e = X^T (I - H) Y = (X^T - X^T H) Y = (X^T - X^T X (X^T X)^{-1} X^T) Y = (X^T - X^T) Y = 0.$$

(2)

$$\hat{Y}^T e = (HY)^T (I - H) Y = Y^T (H - H^2) Y = 0.$$

□

15.1.3 Ordinary least square (OLS): Additional topics

15.1.3.1 *Orthogonal input and successive regression*

This section we cover a special of OLS, in which columns of design matrix X are orthogonal. Under such condition, the parameter estimation for β becomes the surprisingly elegant and geometry interpretation emerges.

Theorem 15.1.6 (multiple linear regression with orthogonal input). *Consider a multiple linear regression problem with n samples can be written as*

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \dots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n(p-1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

with matrix form $Y = X\beta + \epsilon$.

Further assume columns of X ***are orthogonal.***[a] *Then*

- *we have*
$$\hat{\beta}_i = \frac{\langle X_i, Y \rangle}{\langle X_i, X_i \rangle},$$
where X_i *is the column* i *of* X *and* $\langle \cdot, \cdot \rangle$ *is inner product.*
- *Further adding or omitting* ***another orthogonal regressor*** *will not affect other already-determined* $\hat{\beta}_i$.

[a] Note that orthogonality implies y and x are demeaned; that is, $\langle \mathbf{1}, y \rangle = 0 \implies \sum_i y_i = 0, \sum_i x_{i,j} = 0, j = 1, 2, ..., p$.

Proof. From OLS solution [Theorem 15.1.1], we have

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Under the orthogonality assumption, $X^T X$ will be a diagonal matrix such that the component of $\hat{\beta}$ can be written by

$$\hat{\beta}_i = \frac{\langle X_i, Y \rangle}{\langle X_i, X_i \rangle}.$$

After incorporating new orthogonal regressors, the $\hat{\beta}_i$ remain the same.

□

We can use **QR decomposition** to make the input orthogonal to each other; such idea gives the following successive regression method.

Methodology 15.1.1 (successive regression method). *Consider a multiple linear regression problem consisting of column input vectors* $X_1, ..., X_p$ *and output vector* Y.

- *Initialize* $Z_0 = 1$.

- *Regress X_1 on Z_0, and denote Z_1 as the residual vector,*

$$Z_1 = X_1 - \frac{\langle X_1, Z_0 \rangle}{\langle Z_0, Z_0 \rangle} Z_0.$$

- *Similarly, regress X_i on $Z_0, Z_1, ..., Z_{i-1}$ for $i = 2, ..., p$, and denote Z_i as the residual vector,*

$$Z_i = X_i - \sum_{j=0}^{i-1} \frac{\left\langle X_1, Z_j \right\rangle}{\left\langle Z_j, Z_j \right\rangle} Z_j.$$

- *Since $Z_0, Z_1, ..., Z_p$ are orthogonal, then*

$$\hat{\beta}_i = \frac{\langle Z_i, Y \rangle}{\langle Z_i, Z_i \rangle}.$$

15.1.3.2 *Frisch-Waugh-Lovell(FWL) theorem and partial regression*

Frisch-Waugh-Lovell(FWL) theorem concerns how components in OLS estimate of $\hat{\beta}$ are related to each other in a geometrical way.

Theorem 15.1.7 (Frisch-Waugh-Lovell theorem). *Consider a linear regression formulation with standard assumptions*

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon,$$

where $Y \in \mathbb{R}^N, X_1 \in \mathbb{R}^{N\times k_1}, X_2 \in \mathbb{R}^{N\times k_2}, \beta_1 \in \mathbb{R}^{k_1}, \beta_2 \in \mathbb{R}^{k_2}$. We assume X_1, X_2 matrices have full column rank. Let $\hat{\beta}_1, \hat{\beta}_2$ be the least square estimators. It follows that

- $$\hat{\beta}_2 = (X_2^T M_1 X_2)^{-1} X_2 M_1 Y,$$

  where $M_1 = I - H_1, H_1 = X_1(X_1^T X_1)^{-1} X_1^T$.
- *The estimator $\hat{\beta}_2$ can be viewed as the least square estimator from a modified linear regression problem given by*

$$M_1 Y = M_1 X_2 \beta + M_1 \epsilon.$$

- $$Var[\hat{\beta}_2] = \sigma^2 (X_2^T M_1 X_2)^{-1}.$$

Proof. (1) Use Theorem 15.1.1. (2) If we multiply the null orthogonal projector M_1 to both sides of

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

and get

$$M_1 Y = M1 X_2 \beta + M_1 \epsilon.$$

Alternatively, directly apply least square solution to $M_1 Y = M_1 X_2 \beta + M_1 \epsilon$ we also get

$$\begin{aligned}\hat{\beta}_2 &= (X_2^T M_1^T M_1 X_2)^{-1} X_2^T M_1^T M_1 Y \\ &= (X_2^T M_1 X_2)^{-1} X_2^T M_1 Y\end{aligned}$$

where we use $M_1^T = M_1, M_1^T M_1 = M_1$. (3)

$$\begin{aligned}Var\beta_2 &= (X_2^T M_1 X_2)^{-1} X_2^T M_1 Var[Y]((X_2^T M_1 X_2)^{-1} X_2^T M_1 Y)^T \\ &= (X_2^T M_1 X_2)^{-1} X_2^T M_1 \sigma^2 I((X_2^T M_1 X_2)^{-1} X_2^T M_1 Y)^T \\ &= \sigma^2 (X_2^T M_1 X_2)^{-1}.\end{aligned}$$

□

15.1.3.3 *Forecasting analysis with normality assumption*

Lemma 15.1.3 (prediction for simple linear regression). *[3, p. 30]Assume* $\epsilon \sim MN(0, \sigma^2 I)$

- *Given the regressors value* x_0*, the unbiased mean response is defined to be*

  $$\hat{\mu}(y|x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

  such that $E[\hat{\mu}(y|x_0)] = \beta_0 + \beta_1 x_0$*. And if* σ^2 *is known,*

  $$\frac{\hat{\mu}(y|x_0) - E[\hat{\mu}(y|x_0)]}{\sqrt{\sigma^2(1/n + (x_0 - \overline{x})^2/S_{XX})}} \sim N(0,1);$$

  If σ^2 *is unknown,*

  $$\frac{\hat{\mu}(y|x_0) - E[\hat{\mu}(y|x_0)]}{\sqrt{\hat{\sigma}^2(1/n + (x_0 - \overline{x})^2/S_{XX})}} \sim t(n-2).$$

  where

  $$\hat{\sigma}^2 = (SSE)/(n-2).$$

- *The estimation of the response* y_0 *given* x_0 *is*

  $$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0,$$

and

$$Var[y_0 - \hat{y}_0] = \sigma^2(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{XX}}).$$

If σ^2 is known, then

$$\frac{\hat{y}_0 - y_0}{\sqrt{\sigma^2(1 + 1/n + (x_0 - \overline{x})^2/S_{XX})}} \sim N(0,1);$$

If σ^2 is unknown,

$$\frac{\hat{y}_0 - y_0}{\sqrt{\hat{\sigma}^2(1 + 1/n + (x_0 - \overline{x})^2/S_{XX})}} \sim t(n-2).$$

where

$$\hat{\sigma}^2 = (SSE)/(n-2).$$

Proof. (1)(a)Using the results in Theorem 15.1.5, we have

$$\begin{aligned} Var[\hat{\mu}(y|x_0)] &= Var[\hat{\beta}_0 + \hat{\beta}_1 x_0] \\ &= Var[\hat{\beta}_0] + Var[\hat{\beta}_1]x_0^2 + 2x_0 Cov(\hat{\beta}_0, \hat{\beta}_1) \\ &= \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}) + \frac{\sigma^2 x_0^2}{S_{XX}} - 2\frac{\sigma^2 x_0 \overline{x}}{S_{XX}} \\ &= \sigma^2(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{XX}}). \end{aligned}$$

(b) Note that

$$\frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2(n-2)}{n-2},$$

therefore

$$\frac{\hat{\mu}(y|x_0) - E[\hat{\mu}(y|x_0)]}{\sqrt{\sigma^2(1/n + (x_0 - \overline{x})^2/S_{XX})}} / \frac{\hat{\sigma}^2}{\sigma^2} \sim t(n-2).$$

where we also the independence between $\hat{\beta}$ and $\hat{\sigma}^2$ in Theorem 15.1.4. (2) Note that

$$y_0 - \hat{y}_0 = \beta_0 + \beta_1 x_0 + \epsilon - \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Then

$$Var[y_0 - \hat{y}_0] = \sigma^2(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{XX}}).$$

The rest is similar to (1). □

Lemma 15.1.4 (prediction for multiple linear regression). *[3, pp. 94, 99] Assume* $\epsilon \sim MN(0, \sigma^2 I)$

- *Given the regressors value* x_0*, the unbiased mean response is defined to be*

$$\hat{y}_0 = x_0^T \hat{\beta}$$

and

$$Var[\hat{y}_0] = \sigma^2 x_0^T (X^T X)^{-1} x_0$$

If σ^2 *is known, then*

$$\frac{\hat{y}_0 - E[y|x_0]}{\sqrt{\sigma^2 x_0^T (X^T X)^{-1} x_0}} \sim N(0,1);$$

If σ^2 *is unknown,*

$$\frac{\hat{y}_0 - E[y|x_0]}{\sqrt{\hat{\sigma}^2 x_0^T (X^T X)^{-1} x_0}} \sim t(n-p),$$

where

$$\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)}{n-p}.$$

- *The estimation of the response* y_0 *given* x_0 *is*

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0,$$

and If σ^2 *is known, then*

$$\frac{\hat{y}_0 - y_0}{\sqrt{\sigma(1 + x_0^T (X^T X)^{-1} x_0)}} \sim N(0,1);$$

If σ^2 *is unknown,*

$$\frac{\hat{y}_0 - y_0}{\sqrt{\hat{\sigma}^2(1 + x_0^T (X^T X)^{-1} x_0)}} \sim t(n-p).$$

Proof. Note that

$$Var[\hat{y}_0] = Var[x_0^T \hat{\beta}] = x_0^T Cov(\hat{\beta}) x_0 = \sigma^2 x_0^T (X^T X)^{-1} x_0.$$

The rest is similar to Lemma 15.1.3. □

15.1.4 Hypothesis testing and analysis of variance

notations

- $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$, where $\overline{y}$ is the sample mean of y values.
- $SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$, where $\hat{y} = \hat{\beta}x_i$ is the predicted value of y based on observation x_i.
- $SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$.
- $MSE = SSE/(n-p), MSR = SSR/(p-1)$
- $S_{XX} = \sum_i (x_i - \overline{x})^2$, where $\overline{x}$ is the sample mean of x values.

15.1.4.1 *Distribution of coefficients*

Given observations of outcome variables and predictor variables of interest, how can I be sure whether there is a linear relationship between them. We can always carry out OLS procedure and yield some non-zero model coefficients. Because the exist of disturbance, a non-zero coefficient will probably be a false-position result. In this section, we will establish a formal hypothesis framework to quantify how significant a linear relationship exist among outcome and predictor variable.

To perform hypothesis testing, we need to first quantify the probability distribution of fitted parameters.

Lemma 15.1.5 (distribution of linear function of coefficients). *Consider a p-degree ($p = 2$ corresponds to simple linear regression) linear regression problem. Let β be the true parameter and $\hat{\beta}$ be the OLS estimator. Assume $\epsilon \sim MN(0, \sigma^2 I)$. Let w be a p-dimensional constant vector. Then*

- *If σ is known, then*
$$w^T\hat{\beta} \sim N(w^T\beta, \sigma^2 w^T X^T X w).$$
- *If σ is unknown, then*
$$\frac{w^T\hat{\beta} - w^T\beta}{\sqrt{w^T(X^TX)^{-1}w\hat{\sigma}^2}}, \hat{\sigma}^2 = \frac{SSE}{n-p}$$
will follow a Student t distribution with $n - K$ degree.

Proof. When $\epsilon \sim MN(0, \sigma^2 I)$, then $Cov(\hat{\beta}) = Cov(HY) = \sigma^2(X^TX)^{-1}$ [Lemma 15.1.1]. Since $\hat{\beta}$ is the affine transformation of multivariate Gaussian Y, $\hat{\beta}$ is also multivariate Gaussian. Therefore the OLS estimator $\hat{\beta}$ has distribution
$$\hat{\beta} \sim MN(\beta, \sigma^2 X^T X).$$
Therefore,
$$w^T\hat{\beta} \sim N(w^T\beta, \sigma^2 w^T X^T X w).$$

From Theorem 15.1.4, we know that $SSE/\sigma^2 \sim \chi^2(n-p)$ and SSE is independent of $\hat{\beta}$. Therefore, the quantity

$$\frac{\frac{w^T\hat{\beta}-w^T\beta}{\sqrt{\sigma^2 w^T X^T X w}}}{\sqrt{SSE/(\sigma^2(n-K))}}$$

will follow a Student t distribution with $n-K$ degree.

□

Theorem 15.1.8 (distribution of coefficients in linear regression). *[3, p. 93] Assume $\epsilon \sim MN(0, \sigma^2 I)$. Then we have*

- *For multiple linear regression with σ known,*

$$\hat{\beta} \sim MN(\beta, \sigma^2 X^T X).$$

- *If σ^2 is known, then*

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 C_{jj}}} \sim N(0,1), j = 0, 1, ..., p-1;$$

  if σ^2 is unknown, then

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t(n-p), j = 0, 1, ..., p-1,$$

  where C_{jj} is the j diagonal element of $(X^T X)^{-1}$, and

$$\hat{\sigma}^2 = \frac{SSE}{n-p}.$$

- *For simple linear regression with σ known,*

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{XX}),$$

  where $S_{XX} = \sum_{i=1}^n (x_i - \overline{x})^2$.
- *For simple linear regression with σ unknown,*

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2/S_{XX}}} = \sqrt{(n-2)S_{XX}}\frac{\hat{\beta}_1 - \beta_1}{\sqrt{SSE}} =\sim t(n-2),$$

where $t(n-2)$ is the standard t distribution with $n-2$ degree of freedom, $E[\hat{\beta}_1] = \beta_1$ and

$$\hat{\sigma}^2 = \frac{SSE}{n-2}.$$

Proof. (1)(2) Use above lemma Lemma 15.1.5 and set w to be a zero vector with only j entry being 1. (3) Note that

$$\hat{\beta}_i / \sqrt{\sigma^2 C_{ii}} \sim N(0,1),$$

(4) Note that

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2).$$

therefore

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / S_{XX}}} / \sqrt{\frac{SSE}{\sigma^2(n-2)}} = \sqrt{(n-2)S_{XX}} \frac{\hat{\beta}_1 - \beta_1}{SSE} \sim t(n-2)$$

based of t distribution definition [Definition 12.1.24]. □

15.1.4.2 *t test and normality test of single coefficients*

Under normality assumption of the disturbance, fitted coefficielts in OLS follow Gaussian distributions and t distribution. Exploiting these results, we can design z or t tests for the fitted coefficients as follows.

Methodology 15.1.2 (t-Test of regression slope for simple linear regression). *[4, p. 261] **Assume σ is unknown.** Assume $\epsilon \sim MN(0, \sigma^2 I)$. Consider the hypothesis that the slope of a regression line :*

$$H_0 : \beta_1 = \beta_{10} (\ often\ \beta_{10} = 0)$$
$$H_1 : \beta_1 \neq \beta_{10}$$

And t-test statistic is

$$t_0 = \sqrt{(n-2)S_{XX}} \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{SSE}}$$

is a Student t-distribution with $n-2$ degrees of freedom. H_0 is rejected when

$$|t_0| \geq t_{\alpha/2},$$

where α is the confidence level, and

$$P(|T| > t_{\alpha/2}) = \alpha . T \sim t(n-2).$$

Proof. Theorem 15.1.8. □

Methodology 15.1.3 (t-Test of single coefficient for multiple linear regression). *[4, p. 261] **Assume σ is unknown.** Assume $\epsilon \sim MN(0, \sigma^2 I)$. Consider the hypothesis that the slope of a regression line :*

$$H_0 : \beta_1 = \beta_{10} (\ often\ \beta_{10} = 0)$$
$$H_1 : \beta_1 \neq \beta_{10}$$

And t-test statistic:

$$t_0 = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$$

where C_{jj} is the j diagonal element of $(X^T X)^{-1}$, and

$$\hat{\sigma}^2 = \frac{SSE}{n-p},$$

is a Student t-distribution with $n - p$ degrees of freedom.

H_0 is rejected when

$$|t_0| \geq t_{\alpha/2},$$

where α is the confidence level, and

$$P(|T| > t_{\alpha/2}) = \alpha . T \sim t(n-p).$$

Proof. Theorem 15.1.8. □

Methodology 15.1.4 (normality-Test of regression slope for simple linear regression). *[4, p. 261] **Assume σ is known.** Assume $\epsilon \sim MN(0, \sigma^2 I)$. Consider the hypothesis that the slope of a regression line :*

$$H_0 : \beta_1 = \beta_{10} (\ often\ \beta_{10} = 0)$$
$$H_1 : \beta_1 \neq \beta_{10}$$

And the normality-test statistic is

$$N_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2 / S_{XX}}}$$

is a standard normal distribution $N(0, 1)$. H_0 is rejected when

$$|N_0| \geq z_{\alpha/2},$$

where α is the confidence level, and

$$P(|Z| > z_{\alpha/2}) = \alpha . Z \sim N(0,1).$$

Proof. Theorem 15.1.8. □

Methodology 15.1.5 (normality-Test of single coefficient for multiple linear regression). *[4, p. 261]* ***Assume σ is known.*** *Assume $\epsilon \sim MN(0, \sigma^2 I)$. Consider the hypothesis that the slope of a regression line :*

$$H_0 : \beta_1 = \beta_{10} (\ often\ \beta_{10} = 0)$$
$$H_1 : \beta_1 \neq \beta_{10}$$

And the normality test statistic is

$$N_0 = \frac{\hat{\beta}_j - \beta_{10}}{\sqrt{\sigma^2 / C_{jj}}}$$

is a standard normal distribution $N(0,1)$, where C_{jj} is the j diagonal element of $(X^T X)^{-1}$.

H_0 is rejected when

$$|N_0| \geq z_{\alpha/2},$$

where α is the confidence level, and

$$P(|Z| > z_{\alpha/2}) = \alpha . Z \sim N(0,1).$$

Proof. Theorem 15.1.8. □

15.1.4.3 *F lack-of-fit test*

Previous t test allows examination of the significance of each predictor variable in explaining the outcome variable. On the other hand, we can also check the significance of all predictor variables via the following hypothesis:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0$$
$$H_1 : \text{at least one } \beta_j \neq 0$$

We can also examine more general hypothesis via the following.

Lemma 15.1.6 (lack-of-fit in F-Test of multiple regression slope). *[3, p. 80][2, p. 139]* ***Assume σ is unknown.*** *The hypothesis that if there is a specified linear relationship between response y and any of the regressor variables$(x_1, x_2, ..., x_{p-1})$ is given by*[a]*:*

$$H_0 : \beta_0 = \beta_0^0, \beta_1 = \beta_1^0, \beta_2 = \beta_2^0, ...\beta_{p-1} = \beta_{p-1}^0 \ (or \beta = \beta^0)$$
$$H_1 : \textit{at least one } \beta_j \neq \beta_j^0$$

We have

- *The F-test statistic is*

$$F_0 = \frac{(\hat{\beta} - \beta^0)^T X^T X (\hat{\beta} - \beta^0)/p - 1}{SSE/(n-p)},$$

  and it has a F-distribution with $(p, n-p)$ degrees of freedom if H_0 is true.
- *If the hypothesis is given by*

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0$$
$$H_1 : \textit{at least one } \beta_j \neq 0$$

  where we leave β_0 unspecified, then the F-test statistic is

$$F_0 = \frac{SSR/p - 1}{SSE/(n-p)},$$

  and it has a F-distribution with $(p-1, n-p)$ degrees of freedom if H_0 is true.

[a] we can assume y and x are demeaned.

Proof. Suppose H_0 is true, then $\hat{\beta} - \beta_0 = \hat{\beta} - \beta = (X^T X)^{-1} X^T \epsilon$. Therefore,

$$(\hat{\beta} - \beta_0)^T X^T X (\hat{\beta} - \beta_0) = \epsilon^T H \epsilon.$$

Since H has rank p, $\epsilon^T H \epsilon \sim \chi^2(p)$ based on Chi-square decomposition theorem [Theorem 12.4.1]. Also from Theorem 15.1.4, we know that,

$$SSE \sim \chi^2(n-p).$$

Finally, use the F distribution definition. □

Lemma 15.1.7 (F-Test of simple regression slope). ***Assume σ is unknown.*** *The hypothesis that the slope of a regression line a constant, β_{10}:*

$$H_0 : \beta_1 = \beta_{10} = 0$$
$$H_1 : \beta_1 \neq \beta_{10}$$

We have:
(1) The F-test statistic is

$$F_0 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})}{SSE/(n-2)} = \frac{SSR}{SSE/(n-2)}$$

and it has a F-distribution with $(1, n-2)$ degrees of freedom under hypothesis H_0.

Proof. (1) From Theorem 15.1.4, we know that, under null hypothesis, $SSR \sim \chi^2(1), SSE \sim \chi^2(n-2)$.

□

Remark 15.1.6 (Interpretation and hypothesis testing procedure)**.**

- If H_0 is true, then $\beta_1 = 0$, then $F_0 \approx 1$.
- If $\beta_1 \neq 0$, F_0 tends to have large values, then H_0 should be rejected.

15.1.4.4 χ^2 test for variance

Under normality assumption of the disturbance, residuals in OLS follow χ^2 distribution parameterized with σ. This leads us to design tests for the estimated variance.

Methodology 15.1.6 (χ^2-Test of variance). *[4, p. 261]* ***Assume σ is unknown.*** *Assume $\epsilon \sim MN(0, \sigma^2 I)$. Consider the hypothesis that the slope of a p-degree linear regression problem:*

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_1 : \sigma^2 \neq \sigma_0^2$$

And χ^2-test statistic is

$$X = \frac{SSE}{\sigma_0^2}$$

is a χ^2 with $n - p$ degrees of freedom.

H_0 is rejected when $X \geq k_1$ or $X \leq k_2$, where α is the confidence level, and

$$P(M > k_1) = \alpha/2, P(M \leq k_2) = \alpha/2. M \sim \chi^2(n-p).$$

Proof. Theorem 15.1.4. □

15.1.5 Maximum likelihood method with normality assumption

Model parameter can also be carried out in the maximum likelihood framework, in which we need to assume probability distribution of disturbance. Under normality assumption of the disturbance, we have the following.

Theorem 15.1.9 (maximum likelihood estimation of parameters in multiple linear regression). *The likelihood function for the multiple linear regression problem is given by*

$$L(\beta, \sigma^2) = \prod_{i=1}^{N} N(y_i; x_i, \beta, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp(-\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)),$$

and the logarithm of this function is

$$l(\beta, \sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta).$$

We have the following derivatives

- $$\frac{\partial l}{\partial \beta} = \frac{1}{\sigma^2}(Y - X\beta)^T X.$$
- $$\frac{\partial l}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)^T(Y - X\beta).$$
- $$\frac{\partial^2 l}{\partial \beta \partial \beta^T} = -\frac{1}{\sigma^2}X^T X.$$
- $$\frac{\partial^2 l}{\partial \sigma^2 \partial \sigma^2} = \frac{N}{2\sigma^4} - \frac{1}{\sigma^6}(Y - X\beta)^T(Y - X\beta).$$
- $$\frac{\partial^2 l}{\partial \beta \partial \sigma^2} = -\frac{1}{\sigma^4}X^T(Y - X\beta).$$

The first order condition gives the following MLE

- $$\hat{\beta} = (X^T X)^{-1} X^T Y,$$
- $$\hat{\sigma}^2 = \frac{1}{N}(Y - X\hat{\beta})^T (Y - X\hat{\beta})$$

Proof. (1) Note that

$$\frac{\partial l}{\partial \beta} = 0 \implies (Y - X\beta)^T X = 0 \implies X^T Y = X^T X\beta.$$

(2)

$$\frac{\partial l}{\partial \sigma^2} = 0 \implies -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)^T (Y - X\beta) = 0 \implies \hat{\sigma}^2 = \frac{1}{N}(Y - X\hat{\beta})^T (Y - X\hat{\beta}).$$

□

Corollary 15.1.9.1 (maximum likelihood estimation of parameters in simple linear regression). *The likelihood function for the multiple linear regression problem is given by*

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{N} N(y_i; x_i, \beta_0, \beta_1, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2),$$

and the logarithm of this function is

$$l(\beta_0, \beta_1, \sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2.$$

We have the following derivatives

- $$\frac{\partial l}{\partial \beta_1} = -2\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)x_i.$$
- $$\frac{\partial l}{\partial \beta_0} = -2\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i).$$
- $$\frac{\partial l}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2$$

The first order condition gives the following MLE

- $$\hat{\beta}_1 = \frac{\sum_{i=1}^N (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^N (x_i - \overline{x})},$$
- $$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x},$$
- $$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

Remark 15.1.7 (biased variance estimator). Note that the unbiased variance estimator is given by [Theorem 15.1.4]

$$\hat{\sigma}^2 = \frac{1}{N-p}(Y - X\hat{\beta})^T (Y - X\hat{\beta})$$

therefore the MLE variance estimator is biased.

15.1.6 Asymptotic properties of least square solutions

15.1.6.1 *Asymptotic properties of standard OLS*

In a standard linear regression with assumptions A1-A5 in Assumption 15.1 hold and finite samples, the least square estimator is the best linear unibiased estimator based on the Gauss-Markov theorem [Theorem 15.1.3]. If we further assume the error/noise term has normal distribution (A6 in Assumption 15.1), then we derive various hypothesis test statistics, such as t-test and F-test.

The full set of assumptions in Assumption 15.1 can be difficult to satisfy in the real world; however, in the large sample limit, we can loose some assumptions (e.g., the normal distribution assumption) can still achieve nice properties (e.g., deriving t and F statistics) using central limit theorem.

Theorem 15.1.10 (consistence of standard OLS). *[1, p. 64] Consider a multiple linear regression under standard assumption [Assumption 15.1] except for the normality assumption. Also make the additional assumption*

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} X^T X = 0.$$

Then the least square estimator based on n samples [Theorem 15.1.1] given by

$$\hat{\beta}_n = (X^T X)^{-1} X^T Y$$

has the following properties:

- *(Consistence) $\hat{\beta}_n$ is a consistent estimator β*

$$\operatorname{plim} \hat{\beta}_n = \beta.$$

- *$\hat{\sigma}_n^2 = SSE/(n-p)$ is a consistent estimator of σ^2.*
- *(Asymptotic normality) $\hat{\beta}_n$ is converging to a normal distribution*

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{D} MN(0, \sigma^2 Q^{-1}).$$

Proof. (1) Note that

$$\operatorname*{plim}_{n\to\infty} \hat{\beta}_n = \operatorname*{plim}_{n\to\infty} \beta + (\frac{X^T X}{n})^{-1}(\frac{1}{n} X^T u) = \beta + Q^{-1} \cdot 0$$

because $\operatorname{plim}_{n\to\infty} \frac{1}{n} X^T u = 0$ since the sample mean $\frac{1}{n} X^T u$ will converge in probability to its mean, which is zero. Note that

$$Var[\hat{\beta}] = \sigma^2 (\sum_{i=1}^{N} X_i X_i^T)^{-1} = \sigma^2 (NQ)^{-1}.$$

(2) Note that $\hat{\sigma}_n^2$ is an unbiased estimator and its variance will converge to zero . Then use convergence in mean square (from (1)) implies convergence in probability [Theorem 11.10.2 and Theorem 15.1.1]. (3) (informal) It is clear that $\operatorname{plim}_{n\to\infty}(\hat{\beta}_n - \beta) = 0$ because of the consistence of $\hat{\beta}_n$. The variance of $\sqrt{n}(\hat{\beta}_n - \beta)$ is

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_n - \beta)\sqrt{n}(\hat{\beta}_n - \beta)^T &= (\frac{1}{n} X^T X)^{-1}(\frac{1}{\sqrt{n}} X^T u)(\frac{1}{\sqrt{n}} X^T u)^T (\frac{1}{n} X^T X)^{-1} \\
&= (\frac{1}{n} X^T X)^{-1}(\frac{1}{n} X^T u u^T X)(\frac{1}{n} X^T X)^{-1} \\
&= (\frac{1}{n} X^T X)^{-1}(\frac{1}{n} X^T u u^T X)(\frac{1}{n} X^T X)^{-1} \\
&\to Q^{-1}(\frac{\sigma^2}{n} X^T X) Q^{-1} \\
&= \sigma^2 Q^{-1}
\end{aligned}$$

□

Remark 15.1.8 (interpretation). In the large sample size limit, the OLS estimator is not only unbiased but also consistent, indicating that the estimator variance will decrease to zero.

15.1.6.2 *Asymptotic efficiency of standard OLS*

We have just showed that MLE is an asymptotically efficient estimator. Does OLS estimator has similar properties?

If the noises are normally distributed, then the least squares estimator is also the maximum likelihood estimator (MLE). By virtue of being an MLE [Theorem 13.2.6], least squares estimator is **consistent, asymptotically normal and asymptotically efficient.** However, if the noise terms are not normal, then MLE and least square estimator are usually not the same. **If MLE is not a linear estimator, MLE is generally more efficient than least square estimator, at least asymptotically.**

The Gauss–Markov theorem claims that least square estimator is the most efficient linear unbiased estimator; The Gauss–Markov theorem is a finite sample theorem requires no assumption of normality but restricts estimators to be linear and unbiased. On the hand, asymptotic properties of MLE requires no normality neither and broaden the class of estimators beyond linear and unbiased estimators.

15.1.7 Partial and multiple correlation

15.1.7.1 *Multiple correlation coefficient, R^2*

R^2, known as coefficients of determination or coefficients of multiple correlation, is the first metric we check if a model is a good fit to the data. Intuitively, R^2 captures the ratio of the explained variations of the outcome variable over its the total variations in the outcome variable. In this section, we will look into R^2 in depth and draw the connection between R^2 and correlations between outcome and predictor variables.

Definition 15.1.4 (coefficient of determination, coefficient of multiple correlation). *The **coefficient of determination** (or **coefficient of multiple correlation**) R^2 is defined as*

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})}{\sum_{i=1}^{n}(y_i - \overline{y})} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST},$$

where

$$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2 = Y^T(I - \frac{1}{n}J)Y$$
$$SSR = \sum_{i=1}(\hat{y}_i - \overline{y})^2 = Y^T(H - \frac{1}{n}J)Y$$
$$SSE = \sum_{i=1}(y_i - \hat{y}_i)^2 = Y^T(I - H)Y$$

and we use $SST = SSR + SSE$ from Theorem 15.1.4.

Theorem 15.1.11 (properties of coefficient of multiple correlation). *[2, p. 164]*

- *The **coefficient of multiple correlation** of the linear regression is given by*

$$R^2 = \frac{\hat{\beta}^T X^T X \hat{\beta}}{Y^T(I - \frac{1}{n}J)Y} = \frac{Y^T H Y}{Y^T(I - \frac{1}{n}J)Y};$$

  Or equivalently,

$$R^2 = 1 - \frac{\hat{e}^T\hat{e}}{Y^T(I - \frac{1}{n}J)Y}.$$

  where $\hat{e}$ is the residual given by

$$\hat{e} = Y - X^T\hat{\beta} = (I - H)Y, H = X(X^TX)^{-1}X^T.$$

- *For simple linear regression, we have*

$$R^2 = \hat{\rho}^2,$$

  where $\hat{\rho} = S_{XY}/S_XS_Y$ is the sample correlation coefficient between X and Y, and

$$S_{XX} = \sum_i (x_i - \overline{x})^2, S_{XY} = \sum_i (x_i - \overline{x})(y_i - \overline{y}, S_{YY} = \sum_i (y_i - \overline{y})^2,$$

  and

$$S_X = \sqrt{S_{XX}}, S_Y = \sqrt{S_{YY}}.$$

Proof. (1)

$$\begin{aligned} R^2 &= \frac{\hat{\beta}^T X^T X \hat{\beta}}{Y^T (I - \frac{1}{n} J) Y} \\ &= \frac{Y^T H H Y}{Y^T (I - \frac{1}{n} J) Y} \\ &= \frac{Y^T (I - (I - H)) Y}{Y^T (I - \frac{1}{n} J) Y} \\ &= 1 - \frac{\hat{e}^T \hat{e}}{Y^T (I - \frac{1}{n} J) Y} \end{aligned}$$

where $X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY$. (2)

$$\begin{aligned} R^2 &= 1 - \frac{\hat{e}^T \hat{e}}{Y^T Y} \\ &= 1 - \frac{\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sum_i (y_i - \overline{y})^2} \\ &= 1 - \frac{\sum_i (y_i - \overline{y} - \hat{\beta}_1 (x_i - \overline{x}))^2}{\sum_i (y_i - \overline{y})^2} \\ &= 1 - \frac{S_{YY} - 2 S_{XY} \hat{\beta}_1 - S_{YY} \hat{\beta}_1^2}{S_{YY}} \\ &= \frac{\hat{\beta}_1^2 S_{XX}}{S_{YY}} \\ &= \hat{\rho}^2 \end{aligned}$$

where use the fact [Corollary 15.1.1.1] that

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}, \hat{\beta}_1 = \hat{\rho} \frac{S_Y}{S_X} = \frac{S_{XY}}{S_{XX}}.$$

□

Remark 15.1.9 (interpretations and basic properties).

- R^2 is the proportion of variation in Y explained by the predicator X. Obviously,

$$0 \leq R^2 \leq 1.$$

  - $R^2 = 0$ is the case of $\beta_1 = 0$ in simple regression.
  - $R^2 = 1$ is all variations is explained by X.
  - Large R^2 implies small residual, or good fit.
  - R^2 can be increase by increasing the number of predictors.

15.1.7.2 *Partial correlation coefficient*

Given a fitted model, how could we know which predictor variable plays more important roles in explaining the variation of the outcome variables. In this section we introduce partial correlation coefficient associated which predictor variable. Participial correlation coefficient is derived from the R^2 of OLS when X_h is included and R^2 of LOS when it is excluded. More formally, we have the following.

Definition 15.1.5 (partial correlation coefficients). *[2, p. 173] The partial correlation coefficient r_h for predictor variable X_h is defined by*

$$r_h = 1 - \frac{1 - R^2}{1 - R^2_{-h}},$$

where R^2_h is the coefficent of multiple correlation without using X_h.

Remark 15.1.10 (interpretation).

- $r_h = 0$ means that $R^2_{-h} = R^2$ and X_h is not useful.
- $r_h = 1 \implies R^2_{-h} = 1$, and X_h alone can fully explain Y.
- Larger r_h, more important X_h in explaining Y.

Lemma 15.1.8 (calculating partial correlation coefficients). *[2, p. 173] Consider a standard multiple linear problem. The partial correlation coefficient associated with first regressor X_1 is given by*

- $$R^2_1 = 1 - \frac{Y^T N Y}{Y^T Y},$$
  where X is decomposed by $X = [X_1\ W], N = I - W(W^T W)^{-1} W^T$.
- $$r_1 = \frac{Y^T N X_1}{\sqrt{(Y^T N y)(X_1^T N X_1)}}.$$

Proof. Note that

$$\frac{1 - R^2}{1 - R^2_1} = \frac{\frac{Y^T M Y}{Y^T Y}}{\frac{Y^T N Y}{Y^T Y}} = \frac{Y^T M Y}{Y^T N Y} = \frac{Y^T (N - \frac{(NX_1)(NX_1)^T}{X_1 N X_1}) Y}{Y^T N Y} = 1 - \frac{(Y^T N X_1)^2}{(Y^T N y)(X_1^T N X_1)} = 1 - r_1^2$$

where we use the fact that $M = X(X^T X)^{-1} X^T = N - \frac{(NX_1)(NX_1)^T}{X_1 N X_1}$ from Corollary 4.5.8.1. □

Remark 15.1.11 (interpretation).

- From the expression of r_1, it is clear that if X_1 lies in the space spanned by the columns of W then $r_1 = 0$, indicating that X_1 is not useful.
- On the other hand, if X_1 has some component lies in the space spanned by the columns of N and X_1 and Y has some level of correlation, then $r_1 > 0$.

15.1.8 Generalized linear regression (GLR)

15.1.8.1 *Linear regression with structural error*

Generalized linear regression (GLR) extends the OLS by allowing rich covariance structure in the disturbance term. Recall that in the OLS and its standard assumptions, we require the statistical properties of the disturbance to have zero mean and covariance matrix $\sigma^2 I$. In GLS, we allow the covariance matrix of error to be $\sigma^2\Sigma$, which turns out to broaden the application scope of OLS significantly. As we will see, GLS can be used to model cases like

- The structure error model covers the heteroscedastic error case where $Var[\epsilon_i|X] \neq Var[\epsilon_j|X], i \neq j$.
- The structure error model also covers the case where consecutive errors are correlated, for example, $Cov[\epsilon_i, \epsilon_{i-1}|X] = \rho$.

Definition 15.1.6 (multiple linear regression model with structural error). *The multiple linear regression model* ***assumes*** *that a random variable Y has a linear dependency on a non-random vector $X \in \mathbb{R}$ given as*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{p-1} X_1 + \epsilon$$

where $\beta_0, \beta_1, ...$ are unknown model parameters, and ϵ is a random variable. Given the observed sample pairs $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n), x \in \mathbb{R}^{p-1}, y \in \mathbb{R}$ as $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \epsilon_i$ and we ***further make the following assumptions on*** *ϵ as*

- $E[\epsilon_i|X] = 0, \forall i$
- $Var(\epsilon|X) = \Sigma, \forall i, j$

If the data is generated with structural error but we use OLS to fit the model, we can still get an unbiased estimator with no longer uniformly minimum error. We ellaborate on this point in the following note.

Note 15.1.1 (properties of ordinary least square estimator under structural error). [1, p. 302] Note that the ordinary least square estimator is given by

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

which is still unbiased, i.e.,

$$E[\hat{\beta}] = E[(X^T X)^{-1} X^T X\beta] = \beta.$$

However, it is no longer the minimum variance estimator because

- (unbiased)
$$E[\hat{\beta}_{OLS}] = \beta.$$
This is because
$$E[\hat{\beta}_{OLS}] = E[(X^T X)^{-1} X^T Y] = E[(X^T X)^{-1} X^T (X\beta + \epsilon)] = \beta,$$
which is still unbiased.
- $$Var[\hat{\beta}_{OLS}] = \sigma^2 (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1}.$$
- (finite-sample inefficient) However, OLS estimator is no longer an efficient estimator (i.e., minimum variance) as pointed out in Theorem 15.1.13.
- Regular t, χ^2, F tests will fail as the normality assumption of ϵ is violated.

15.1.8.2 *Generalized least square solution*

Analogous to OLS solution to ordinary linear regression models, we can also have generalized least square (GLS) solution to GLR by introducing a weighting scheme in the objective function. Recall that in the OLS, we seek model parameters to minimize an objective function given by

$$\min_{\beta} (Y - X\beta)^T (Y - X\beta).$$

In GLS, the modified objective function is given by

$$\min_{\beta} (Y - X\beta)^T \Sigma^{-1} (Y - X\beta).$$

More details of GLS is given as follows.

Theorem 15.1.12 (generalized least square solution: general case). *The multiple linear regression with n samples can be written as*

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \dots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n(p-1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_{p-1} \end{bmatrix}$$

with matrix form

$$Y = X\beta + \epsilon$$

*The **unique minimizer** to the problem*

$$\min_{\beta} J(\beta) = (Y - X\beta)^T \Sigma^{-1} (Y - X\beta)$$

is given as

$$\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y,$$

in particular,

$$\hat{\beta}_0 = \overline{y} - \sum_{i=1}^{p} \hat{\beta}_i \overline{x}$$

$$\hat{y}_i - \overline{y} = \sum_{i=1}^{p} \hat{\beta}_i (x_i - \overline{x})$$

Moreover, we have

- $E[\hat{\beta}] = \beta$.
- *If* $Cov[Y] = \sigma^2 \Sigma$, *then* $Cov[\hat{\beta}] = \sigma^2 (X^T \Sigma X)^{-1}$.*($\hat{\beta}$ is not necessarily normal)*
- *With known* Σ, *one unbiased estimator for* σ^2 *is*

$$\hat{\sigma}^2 = \frac{1}{n-p} (Y - X\beta)^T \Sigma^{-1} (Y - X\beta)$$

Proof. (1) $J(\beta) = Y^T \Sigma^{-1} Y - \beta^T X^T \Sigma^{-1} X\beta - 2\beta^T X^T \Sigma^{-1} Y$. Set $dJ/d\beta = 0$, we can get the result. (2)(a) (unbiasedness)

$$E[\beta] = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} E[Y] = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} X\beta = \beta.$$

(3)(b) (variance)

$$Cov[\beta] = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Cov[Y]((X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1})^T = \sigma^2 (X^T \Sigma^{-1} X)^{-1}.$$

□

Note 15.1.2 (connection to ordinary least square). Let Σ^{-1} be symmetric positive definite, and let $\Sigma^{-1} = U\Lambda U^T$. Let $S^{-1} = U\sqrt{\Lambda}, S^{-1}S^{-T} = \Sigma^{-1}$, then the transformed model

$$S^{-1}Y = S^{-1}X\beta + S^{-1}\epsilon$$
$$Y^* = X^*\beta + \epsilon^*$$

becomes the canonical linear regression model such that

$$E[\epsilon^*] = E[S^{-1}\epsilon] = S^{-1}E[\epsilon] = 0$$

and

$$Var[\epsilon^*] = P^{-1}Var[\epsilon]P^{-T} = S^{-1}\sigma^2\Sigma S^{-T} = \sigma^2 I.$$

Then we can apply the ordinary least square to the transformed data Y^*, X^*.

15.1.8.3 *Gauss-Markov theorem for GLR*

Similarly to OLS (cite), we have a Gauss-Markov theorem that holds for GLS, stating that GLS is the BLUE when there is structural error.

Theorem 15.1.13 (Gauss-Markov theorem for general least square solution, best linear unbiased estimator). *Given the statistical model*

$$Y = X\beta + \epsilon, E[\epsilon] = 0, Cov(\epsilon) = \sigma^2\Sigma$$

*with β being the model parameter, y being the observations, the **uniformly minimum variance estimators among all linear unbiased estimators** is given by*

$$\hat{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y.$$

As a summary, we have

- $E[\hat{\beta}] = \beta$.
- $Cov[\hat{\beta}] = \sigma^2(X^TX)^{-1}$.
- *Furthermore, if ϵ is Gaussian noise, i.e., $\epsilon \sim MN(0, \sigma^2\Sigma)$, then $\hat{\beta}$ is the uniformly minimum variance estimator **among all estimators**.*

Proof. (1) (unbiasedness)

$$\begin{aligned} E[\hat{\beta}] &= E[(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y] \\ &= E[(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}(X\beta+\epsilon)] \\ &= \beta + E[(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\epsilon] \\ &= \beta \end{aligned}$$

where we use the independence between ϵ and X. (2) Let $\theta' = AY$ be any other unbiased linear estimator, and assume $\theta' = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1} + D$ for some matrix D. The unbiasedness requires that

$$E\theta' = \beta \Rightarrow DX = 0.$$

The variance of the estimator is given by

$$\begin{aligned} &E[(\theta'-\theta)(\theta'-\theta)^T] \\ &= E[(D+(X^TX)^{-1}X^T)\epsilon\epsilon^T(D+(X^TX)^{-1}X^T)^T] \\ &= \sigma^2[(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}+D]\Sigma[\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}+A^T] \\ &= \sigma^2(X^T\Sigma^{-1}X)^{-1} + \sigma^2A\Sigma A^T + \sigma^2(X^TV^{-1}X)^{-1}X^TA^T + \sigma^2AX(X^T\Sigma^{-1}X)^{-1} \\ &= \sigma^2(X^T\Sigma^{-1}X)^{-1} + \sigma^2A\Sigma A^T \geq \sigma^2(X^T\Sigma^{-1}X)^{-1} \end{aligned}$$

where the last two terms zero because the unbiasedness requirement of $DX = 0$ and the $\geq$ sign is in the semi-positive matrix sense. (3) Similar to the proof in Theorem 15.1.3. □

15.1.8.4 *Feasible GLS*

In practice, Σ is unknown and also requires estimation. We can first estimate Σ from the residuals of the OLS, and then carry out GLS using estiamted Σ. The procedures are summarized as follows.

Methodology 15.1.7 (feasible GLS). *Consider a linear regression problem with structure error Definition 15.1.6. The two-step procedure to estimate β is*

- *first obtaining estimator Ω from the OLS residuals.*
- *second estimate $\hat{\beta}$ using GLS estimator*

$$\hat{\beta}_{FGLS} = (X^T\hat{\Omega}^{-1}X)^{-1}X^T\hat{\Omega}^{-1}Y.$$

15.1.9 Linear structure in joint distributions

In our previous treatment on linear regression models, we assume observations of non-random predictors $X_1, ..., X_p$ are given. We can also assume predictors are random

variables and we are given their samples, which turns out will not change the previous result.

By assuming random regressors, we can alternatively model the linear relationship in the joint distribution of predictors and outcome variables. The route of treatment can draw connections among linear regression theory, conditional expectation theory, and best linear predictor theory, as we state in the following.

Theorem 15.1.14 (multivariate Gaussian distribution and multiple linear regression). *Suppose $(x_1, x_2, ..., x_p)$ and y are jointly distributed according to multivariate Gaussian distribution with parameter (μ, Σ). Then*

- *the conditional distribution of y given $(x_1, x_2, ..., x_n)$ is given by*

$$\begin{aligned} & f(y|x_1, x_2, ..., x_n) \\ =& \frac{1}{\sqrt{2\pi(\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})}} \exp(-\frac{1}{2}(\frac{(y - \mu_Y - \Sigma_{YX}\Sigma_{XX}^{-1}(x - \mu_X))^2}{(\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})}), \end{aligned}$$

  where we decompose

$$\mu = [\mu_Y^T, \mu_X^T]^T, \Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix},$$

  with $\mu_1 \in \mathbb{R}^k, \mu_2 \in \mathbb{R}^{n-k}$.
- *The conditional mean $Y|X$ are related linearly to X by*

$$E[Y|X] = \mu_y + \Sigma_{YX}\Sigma_{XX}^{-1}(x - \mu_X) = \mu_Y + \beta(x - \mu_X), \beta = \Sigma_{YX}\Sigma_{XX}^{-1}$$

- *And the conditional variance is given by and,*

$$Var[Y|X] = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}.$$

Proof. This is a re-statement of multivariate conditional distribution Theorem 14.1.2 □

Remark 15.1.12 (connections to OLS and the best linear predictor theory)**.**

- In OLS, $\hat{\beta} = (X_d^T X_d)^{-1} X_d^T Y$, where X_d is the data matrix. In here, we have $\beta = \Sigma_{YX}\Sigma_{XX}^{-1}$
- The result is closely related best linear predictor [Theorem 11.8.2].

Corollary 15.1.14.1 (bivariate Gaussian distribution and linear regression). *[3, p. 49] Suppose x and y are jointly distributed according to bivariate distribution with parameter $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ such that*

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}[(\frac{y-\mu_1}{\sigma_1})^2 + (\frac{y-\mu_1}{\sigma_1})^2 - 2\rho(\frac{y-\mu_1}{\sigma_1})(\frac{x-\mu_2}{\sigma_2})]).$$

Then

- *the conditional distribution of y given x is*

$$f(y|x) = \frac{1}{\sqrt{2\pi}\sigma_{1,2}} \exp(-\frac{1}{2}(\frac{(y-\beta_0-\beta_1 x)^2}{\sigma_{1,2}^2})$$

  where

$$\beta_0 = \mu_1 - \mu_2\rho\frac{\sigma_1}{\sigma_2}, \beta_1 = \rho\frac{\sigma_1}{\sigma_2}, \sigma_{1,2}^2 = \sigma_1^2(1-\rho^2), \rho = \frac{\sigma_{1,2}}{\sigma_1\sigma_2}.$$

- *The conditional mean and conditional variance are given by*

$$E[y|x] = \beta_0 + \beta_1 x, Var[y|x] = \sigma_{1,2}^2.$$

As a final note, we discuss the conditions under which our previous results will still hold.

Note 15.1.3. [3, p. 49] Suppose that $X = (X_1, ..., X_p)$ and Y are jointly distributed random variables but the form of this joint distribution is unknown. All of our previous regression results hold if the following conditions are satisfied.

- The conditional distribution of Y given $X = x$ is normal with conditional mean $\beta_0 + \beta_1 x$ and conditional covariant matrix Σ.
- The X is a random variable or a random vector whose probability distribution does not involve β_0, β_1, and Σ.

15.2 Model specification and selection

15.2.1 Model order mis-specification

15.2.1.1 *Omission of relevant regressors*

The first basic case of model order mis-specification is omitting relevant regressors. For example, the ground truth model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$, but we assume the model to be $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.

Clearly, omitting relevant variables will give a model with larger prediction error on the seen samples. In terms of parameter estimator, the estimator is no longer unbiased but has a smaller estimator variance.

Lemma 15.2.1 (least square estimator with omission of relevant regressors). *Suppose the correct model is given by*

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon,$$

where X_1, X_2 are the data matrices. But the assumed model is $y = X_1\beta_1 + \epsilon$. Suppose all other assumptions in Assumption 15.1 hold. Then

- *The least square estimator is biased and given by*

$$\hat{\beta}_1 = \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2.$$

- *Let $\hat{\beta}_1*$ be the least square estimator of coefficient β_1 with correctly specified model. Then*

$$\hat{\beta}_1^* = (X_1^T M_2 X_1)^{-1} X_1 M_2 Y,$$

  where $M_2 = I - H_2, H_2 = X_2(X_2^T X_2)^{-1} X_2^T$.
- *Small model gives a biased estimator with smaller variance; that is,*

$$\sigma^2 (X_1^T X_1)^{-1} = Var[\hat{\beta}_1] \leq Var[\hat{\beta}_1^*] = \sigma^2 (X_1^T M_2 X_1)^{-1}.$$

Proof. (1)

$$\begin{aligned}
\hat{\beta_1} &= (X_1^T X_1)^{-1} X_1^T Y \\
&= (X_1^T X_1)^{-1} X_1^T (X_1\beta_1 + X_2\beta_2 + \epsilon) \\
&= \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2 + (X_1^T X_1)^{-1} X_1^T \epsilon \\
&= \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2
\end{aligned}$$

(2) From Theorem 15.1.7. (3)

$$\begin{aligned}
&Var[\hat{\beta}_1]^{-1} - Var[\hat{\beta}_1^*]^{-1} \\
&= \frac{1}{\sigma^2}(X_1^T X_1 - X_1^T M_2 X_1) \\
&= \frac{1}{\sigma^2}(X_1^T (I - M_2) X_1) \\
&= \frac{1}{\sigma^2}(X_1^T H_2 X_1) \\
&\geq 0
\end{aligned}$$

where it is clear that

$$(X_1^T H_2 X_1) = (X_1^T H_2^T H_2 X_1) = (H_2 X_1)^T (H_2 X_1) \geq 0.$$

□

15.2.1.2 *Inclusion of irrelevant regressors*

The second basic case of model order mis-specification is including irrelevant regressors. For example, the ground truth model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, but we assume the model to be $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$.

Intuitively, including irrelevant variables will give a model with smaller prediction error on the seen samples, as we are optimizing over more parameters. In terms of parameter estimation, the estimator is still unbiased but has a larger estimator variance.

Lemma 15.2.2 (least square estimator with inclusion of relevant regressors). *Suppose the correct model is given by*

$$Y = X_1 \beta_1 + \epsilon,$$

where X_1 is the data matrix. But the assumed model is $Y = X_1 \beta_1 + X_2 \beta_2 + \epsilon$, where the data generation process for X_2 is independent/irrelevant to X_1, Y, ϵ. Suppose all other assumptions in Assumption 15.1 hold. Then

- *The least square estimator is given by*

$$\hat{\beta}_1 = (X_1^T M_2 X_1)^{-1} X_1^T M_2 Y,$$

where $M_2 = I - H_2, H_2 = X_2 (X_2^T X_2)^{-1} X_2^T$. And it is unbiased

$$E[\hat{\beta}_1] = \beta_1.$$

- *The variance of the estimator is larger than the variance of estimator in a correctly specified model. That is*

$$Var[\hat{\beta}_1] = \sigma^2 E[(X_1^T M_2 X_1)^{-1}] \geq \sigma^2((X_1^T X_1)^{-1}.$$

Proof. (1) From Theorem 15.1.7.

$$\begin{aligned} E[\hat{\beta}_1] &= E[(X_1^T M_2 X_1)^{-1} X_1^T M_2 Y] \\ &= E[(X_1^T M_2 X_1)^{-1} X_1^T M_2 (X_1\beta_1 + \epsilon)] \\ &= \beta_1 + 0 = \beta_1 \end{aligned}$$

(2) In a correctly specified model, we have $\hat{\beta}_1^{corr} = (X_1^T X_1)^{-1} X_1^T Y$ and its variance is $\sigma^2(X_1^T X_1)^{-1}$.

In our mis-specified model, we know

$$Var \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \sigma^2 (X^T X)^{-1},$$

where $X = [X_1 \ X_2]$.

By inverting the block matrix $X^T X$ [Lemma A.8.6], we have

$$Var[\hat{\beta}_1] = \sigma^2((X_1^T X_1)^{-1} + B_{12} B_{22}^{-1} B_{21})$$

where

$$B_{12} = (X_1^T X_1)^{-1} X_1^T X_2, B_{21} = X_2^T X_1 (X_1^T X_1)^{-1}$$

and

$$B_{22} = (X_2^T X_2) - X_2^T X_1 (X_1^T X_1)^{-1} X_1^T X_2 = X_2^T (I - M_1) X_2 \geq 0.$$

Because B_{22} is semi-positive positive and so are the cases with B_{22}^{-1} and $B_{12} B_{22}^{-1} B_{21}$.

We have

$$Var[\hat{\beta}_1] = \sigma^2((X_1^T X_1)^{-1} + B_{12} B_{22}^{-1} B_{21}) \geq \sigma^2((X_1^T X_1)^{-1}$$

□

15.2.2 Model selection methods

15.2.2.1 *Adjusted R square method*

As we see subsubsection 15.1.7.1, given a linear model and data, R^2 provides the first glance if the model fits the data well. Given finite-sided data, a model with more predictor variables will increase R^2 even though the model could overfit the data. Therefore, R^2 lacks the ability to check and prevent overfitting. To provide goodness-of-fit and model overfitting gauge at the same time, we can modify R^2 to the following **adjusted** R^2.

Definition 15.2.1 (adjusted R^2). *The **adjusted** R^2 is defined by*

$$R^2_{adj} = 1 - \frac{(1-R^2)(N-1)}{N-K-1},$$

where

- *N is the number of points in the data sample.*
- *K is the number of independent regressors, excluding the constant.*

Remark 15.2.1 (interpretation).

- Increasing number of regressors will increase R^2; however, it might decrease $\mathbb{R}^2_{adj}$.
- We can choose the optimal number of regressors that maximize $\mathbb{R}^2_{adj}$.

15.2.2.2 *F test method*

Using a larger model can always reduce the residual and yield larger R^2. But we expect the marginal gain will gradually become smaller as we continue to increase the size of the model. One characteristic of an overfitting model is on average each prediction variable contribute to a small portion of explained variations. Given two models, we can compare **their average residual reduction by per predictor variable** to examine which model is at lower risk of overfitting, which gives the following F test theory.

Theorem 15.2.1 (F-Test of different linear models). *Consider the null hypothesis that a smaller model is better:*

$$M_0(\textit{small model}, \textit{null hypothesis}) : y = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q, \beta_{q+1} = ... = \beta_{p-1} = 0$$
$$M_1(\textit{large model}, \textit{alternative hypothesis}) : \beta_j \neq 0, \forall j = 1, ..., p.$$

We have:

- $$SSE(M_1) \sim \sigma^2\chi^2(n-p), SSE(M_0) - SSE(M_1) \sim \sigma^2\chi^2(p-q),$$

  where

  $$SSE = \sum_{i=1}(y_i - \hat{y}_i)^2 = Y^T(I-H)Y.$$

- *The F-test statistic is*

  $$F_0 = \frac{(SSE(M_0) - SSE(M_1))/(p-q)}{SSE(M_1)/(n-p)}$$

  is a F-distribution with $(p-q, n-p)$ degrees of freedom.

- *The criterion to reject the null hypothesis is*

  $$F_0 > F_\alpha.$$

Proof. Note that

$$SSE(M_0) - SSE(M_1) = Y^T(I-H_0)Y - Y^T(I-H_1)Y = Y^T(H_1-H_0)Y,$$

where H_0 and H_1 are the hat matrix associated with the model M_0 and M_1. Then we can show that

$$(H_1-H_0)^2 = H_1^2 - H_1H_0 - H_0H_1 + H_0^2 = H_1 - 2H_0 + H_0 = H_1 - H_0,$$

therefore $H_1 - H_0$ is a orthogonal projector with rank equals

$$Tr(H_1-H_0) = Tr(H_1) - Tr(H_0) = p - q.$$

Then, we can use Cochran's theorem [Theorem 12.4.3]. □

Remark 15.2.2 (Decision rule and implications).

- If H_0 is true, then $F_0 \approx 0$. So we will reject F_0 when F_0 exceeds certain critical value(for example, 95% percentile of the F-distribution).
- Using larger model will always reduce the residual; however, if the error reduction is insignificant, the increase p will make F_0 smaller via

  $$\frac{n-p}{p-q},$$

  indicating that smaller model is better. On the other hand, if n is large, then we are more easy to accept larger model.

Example 15.2.1 (testing all parameters are zero). Consider the 3-parameter full linear regression model given by

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i;$$

And consider the following hypothesis

- $M_0 : \beta_1 = \beta_2 = \beta_3 = 0$.
- $M_1 : at\ least\ one\ \beta_j \neq 0, j = 1, 2, 3$.

Then the statistic

$$F = \frac{(SSE(M_0) - SSE(M_1))/(3)}{SSE(M_1)/(n-4)} = \frac{(SST - SSE(M_1))/(3)}{SSE(M_1)/(n-4)} = \frac{(SSR)/(3)}{SSE(M_1)/(n-4)},$$

which recovers results in Lemma 15.1.6.

Example 15.2.2 (testing one slope parameters is zero). Consider the 3-parameter full linear regression model given by

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i;$$

And consider the following hypothesis

- $M_0 : \beta_1 = 0$.
- $M_1 : \beta_1 \neq 0$.

Then the statistic

$$F = \frac{(SSE(M_0) - SSE(M_1))/(1)}{SSE(M_1)/(n-4)}$$

which has F-distribution with $(1, n-4)$ degrees of freedom.

15.2.2.3 *Information criterion methods*

Definition 15.2.2 (Akaike's Infomration Criterion(AIC)). *The Akaike's Infomration Criterion(AIC) is defined as*

$$AIC = \log(\hat{\sigma}_k^2) + \frac{n + 2k}{n}$$

where $\hat{\sigma}_k^2 = SSR/(n-k)$ *and* k *is the number of parameters in the model.*

Definition 15.2.3 (Bayesian information criterion(BIC)). *The Bayesian Infomration Criterion(BIC) [Lemma 15.2.3] is defined as*

$$BIC = \log(\hat{\sigma}_k^2) + \frac{k\log(n)}{n}$$

where $\hat{\sigma}_k^2 = SSR/(n-k)$ and k is the number of parameters in the model.

Remark 15.2.3 (decision rule).

- We will choose the model that maximize the BIC value.
- If n is large, then we tend to choose larger model.
- When n is fixed, increase k will decrease the first term but increase the second term.

15.2.2.4 *Bayesian information criterion (BIC)*

Definition 15.2.4 (Bayesian information criterion(BIC)). *The BIC is formally defined by*

$$BIC = k\ln(n) - 2\ln(\hat{L}),$$

where

- *$\hat{L}$ is the maximized value of the likelihood function of the model M, i.e., $\hat{L} = p(x|\hat{\theta}, M)$, where $\hat{\theta}$ are the parameter values that maximize the likelihood function and x is the observed data set.*
- *n is the number of the observation data*
- *k is the number of parameters estimated by the model.*

Lemma 15.2.3 (BIC for multiple linear regression). *The BIC for a multiple linear regression model is given by*

$$BIC(M) = k\log(n) + n\log(RSS/n),$$

where

$$RSS = \sum_{i=1}^{n}(Y_i - \hat{\beta}^T X_i)^2$$

Proof. In the linear regression model, we have

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon = f(X) + \epsilon,$$

where ϵ is normal variable with zero mean and a variance of σ. The likelihood function for parameter $\beta_0, ..., \beta_n, \sigma$ is given by

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(Y_i - f(X_i))^2}{2\sigma^2}).$$

The maximum likelihood will be achieved [Theorem 15.1.9] at

$$\hat{\sigma}^2 = \frac{1}{n} RSS,$$

where $RSS = \sum_{i=1}^{n} (Y_i - f(X_i))^2$ such that

$$\ln(\hat{L}) = -\frac{n}{2} \ln(RSS/n) - n/2 - \frac{n}{2} \ln 2\pi.$$

Then $-2\ln(\hat{L}) = n \log(RSS/n) + const.$ □

15.2.3 Test for structure change

In economical or social science studies, a time-series model can often contain a structural break, due to a change in policy or the occurrence of major economic event. For example, when there is a structure break, the model

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t,$$

holds in the time period considered.

Suppose there is a structural break at time t_1, then before t_1, the model is given by

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t;$$

after the structural break, the model is given by

$$y_t = \delta_0 + \delta_1 x_t + \epsilon_t,$$

where $\beta_0 \neq \delta_0$ and $\beta_1 \neq \delta_1$.

The insight is that after a structural break, an alternative model that account for the changing dynamics could explain the data more effective. Similar to F test to compare the effectiveness of two models [Theorem 15.2.1], we have following F test for structural break.

Methodology 15.2.1 (F test for structural break).

- *Let M_0 be the null hypothesis model if there is no structural break given by*

$$M_0 : y_t = \beta_0 + \beta_1^T x_t + \epsilon_t, \beta_1 \in \mathbb{R}^{p-1}$$

Let t_1 be the structural breaking time point. Let M_1 be the alternative model given by

$$M_1 : y_t = \beta_0 + \beta_1^T x_t + \alpha_0 \mathbf{1}(t > t_1) + \alpha_1 x_t \mathbf{1}(t > t_1)\epsilon_t.$$

- *The F-test statistic [Theorem 15.2.1] is*

$$F_0 = \frac{(SSE(M_0) - SSE(M_1))/(p-q)}{SSE(M_1)/(n-p)}$$

is a F-distribution with $(p, n-2p)$ degrees of freedom.

- *The criterion to reject the null hypothesis is*

$$F_0 > F_\alpha.$$

15.3 Linear regression analysis: diagnostics & solutions

15.3.1 Multi-collinearity

15.3.1.1 *Detection and characterization*

Multi-collinearity, or collinearity is the phenomenon that some predictors are linear combinations of the other predictors, or some predictors are highly correlated. As a direct consequence, $X^T X$ is singular or has a large conditional number. More specifically, it can cause the following detrimental effects:

- A unique $\hat{\beta}$ cannot be found.
- In the case of near singularity, $Var[\hat{\beta}] = \sigma^2 (X^T X)^{-1}$ is huge.

The multi-collinearity can be detected either through a coefficient linear regression approach and through variance inflation factor method, as we will introduce later. To remedy multi-collinearity issue, we can either remove the regressors causing multi-collinearity or we can use principal component linear regression method, as we explore in the following sections.

Remark 15.3.1 (connection with machine learning training procedure).

- In machine learning, if the training examples are highly linear dependent, then the variance of the model is big, leading to large test error.
- Correlated examples do not provide too much new information, and therefore training examples should be uncorrelated examples.

15.3.1.2 *Regressor linear regression and variance inflation factor*

Because collinearity is caused by the high correlation among regressors, one needs to find out the regressor that is most correlated with the rest and at same time plays insignificant role in explaining the variation of the outcome variable. One can regress regressors of interest on the rest of regressors and remove those with high R^2, as stated in the following. Note that we prefer R^2 instead of fitted coefficients to gauge correlation since fitted coefficients are scale dependent.

To what extent a high R^2 is alarming can be captured by the concept of **variance inflation factor (VIF)**, which is defined in the following way.

Definition 15.3.1 (variance inflation factor, VIF). *[3, p. 335] In multiple linear regression, the **variance inflation factor**(VIF) associated with predictor X_i is defined by*

$$VIF_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of determination from regression of X_i on the rest of other predictors.

The calculation of VIF and its implications relies on the following Lemma.

Lemma 15.3.1 (basic properties and relations on VIF). *Let R_1^2 be the coefficient of determination from regression of X_1 on the rest of other predicators. It follows that*

- $$R_1^2 = \frac{X_1^T H_{-1} X_1}{X_1^T X_1}.$$
- $$Var[\hat{\beta}_1] = \frac{\sigma^2}{X_1^T X_1} VIF_1 = \frac{\sigma^2}{X_1^T X_1} \frac{1}{1 - R_1^2}.$$

Proof. (1) From Theorem 15.1.11. (2) From Theorem 15.1.1, we have

$$\hat{\beta}_1 = \frac{X_1^T (I - H_{-1}) Y}{X_1^T (I - H_{-1}) X_1},$$

where $H_{-1} = X_{-1}(X_{-1}^T X_{-1})^{-1} X_{-1}^T$, X_{-1} is the matrix without column i. Then

$$\begin{aligned}
Var[\hat{\beta}_1] &= \frac{X_1^T (I - H_{-1}) Y Y^T (I - H_{-1}) X_1}{(X_1^T (I - H_{-1}) X_1)^2} \\
&= \frac{X_1^T (I - H_{-1}) X_1 \sigma^2}{(X_1^T (I - H_{-1}) X_1)^2} \\
&= \frac{\sigma^2}{X_1^T (I - H_{-1}) X_1} \\
&= \frac{\sigma^2}{X_1^T X_1 - X_1^T X_1 R_1^2} \\
&= \frac{\sigma^2}{X_1^T X_1 (1 - R_1^2)} \\
&= \frac{\sigma^2}{X_1^T X_1} VIF_1
\end{aligned}$$

□

Example 15.3.1. [3, p. 326]

$$y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

and the least-square normal equation are

$$(X^T X)\hat{\beta} = X^T y,$$

or equivalently

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

where r_{12} is the correlation between x_1 and x_2, and r_{1y} is the correlation between x_1 and y.

Then the inverse $(X^T X)$ is

$$(X^T X)^{-1} = \begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix},$$

and the estimates of the regression coefficients are

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12} r_{2y}}{1 - r_{12}^2}, \hat{\beta}_2 = \frac{r_{2y} - r_{12} r_{1y}}{1 - r_{12}^2}.$$

The variance of estimator $\hat{\beta}_1$ and $\hat{\beta}_2$ are

$$Var[\hat{\beta}_i] = \sigma^2 \frac{1}{1 - r_{12}^2}.$$

If there is a strong multi-collinearity between x_1 and x_2, then the correlation coefficient r_{12} will be close to 1.

As the consequence,

- the variance of estimator $\hat{\beta}$ is large; in other words, the result will be unstable, different sample input will give very different $\hat{\beta}$.
- the magnitude of estimator $\hat{\beta}$ also tends to be large.

Now we are in a position to give the VIF method for multi-collinearity detection.

Methodology 15.3.1 (VIF method to detect multi-collinearity).

- *For every regressor X_i, regress it on the rest of regressors.*
- *Compute VIF_i. In general, $VIF_i > 5$ is considered bad.*

15.3.1.3 *Principal component linear regression (PCLR)*

Principal component analysis (PCA) is one commonly tool for dimensionality reduction and data compression. In the linear regression setting, we can project the original regressor observations into a small set of new regressor observations. The new set of regression observations are guaranteed uncorrelated by PCA, thus removing the lead cause for multi-collinearity.

The final method is known as **principal component linear regression (PCLR)**.

Methodology 15.3.2 (principal component linear regression, PCLR). *Suppose the eigendecomposition of X^TX is given by*

$$X^TX = U\Lambda U^T.$$

We take the k eigenvectors with the largest eigenvalues and form matrix

$$U_k = [u_1, u_2, ..., u_k].$$

The transformed regressors are $Z_k = XU_k$ and the linear model in the transformed regressors is given by

$$y = Z_k\alpha + \epsilon,$$

where $\alpha \in \mathbb{R}^k$.

15.3.2 Rank deficiency and rigid regression

If the number of predictors is exceeding the number of observations, the OLS solution $\hat{\beta} = (X^TX)^{-1}X^TY$ will break down since the matrix X does not have full column rank and X^TX has rank deficiency and is not invertible [Lemma 4.4.3].

One approach to rank deficiency is to modify the objective function in OLS to include a norm L^2 penalty term on the fitted coefficients. The resulting linear regression is known as rigid regression.

Definition 15.3.2 (ridge regression). *Consider **centered** data in the linear regression model. The ridge regression is to estimate $\hat{\beta}_{ridge}$ by minimizing*

$$\min_{\beta} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2.$$

Theorem 15.3.1 (solution and properties in ridge regression).

- *The solution to the ridge regression is*

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y.$$

- *The ridge estimator for β is **biased***

$$E[\hat{\beta}_{ridge}] = (X^T X + \lambda I)^{-1} X^T X \beta \neq \beta,$$

  whereas in the ordinary linear regression $\hat{\beta}_{OLS} = \beta$. Particularly if we denote the eigen-decomposition $X^T X = U \Sigma U^T$, then

$$E[\hat{\beta}_{ridge}] = U(\Sigma + \lambda I)^{-1} \lambda I \beta.$$

- *The covariance of the and*

$$Cov[\hat{\beta}_{ridge}] = \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}$$

Proof. (1)(2) Direct optimization. □

Remark 15.3.2 (requirement for data preprocessing).

- For ridge regression, we need to preprocess that data, because the shrinkage effect depends on the scale of the predicators. For canonical linear regression, we do not need to preprocess the data in prediction(the $X\hat{\beta} = X(X^T X)^{-1} X^T Y$ from preprocessed data and original data make no difference. Note that scaling amounts to times a diagonal matrix).
- The optimal λ is usually selected via cross-validation algorithms [see subsection 23.2.1].

15.3.3 Heteroskedasticity

15.3.3.1 *Test for heteroskedasticity*

In linear regression, **heteroskedasticity** is the situation where the conditional variance of noise ϵ_i is not a constant across different observed x_i, but has a dependence on x_i, as we show in Figure 15.3.1. With heteroskedasticity, OLS estimator is still unbiased. However, it is no longer efficient and t, F tests are no longer valid because of their reliance on normality assumption, as we discussed in 15.1.1.

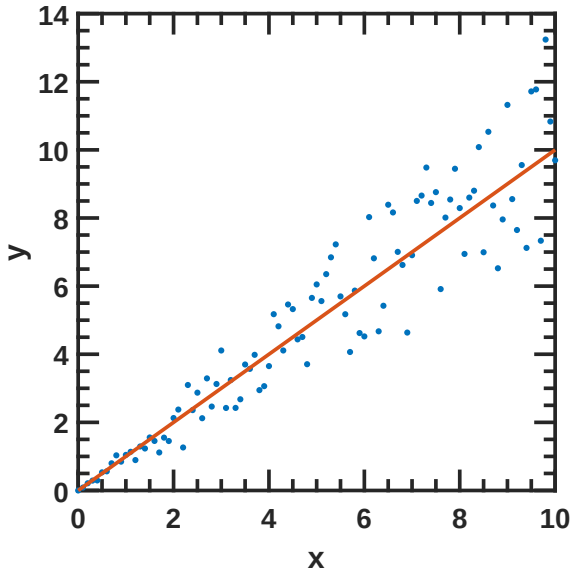


Figure 15.3.1: Demo of heteroskedasticity in linear regression. The noises are larger at larger x values.

Designing a heteroskedasticity test can harness the following intuitions: In the standard linear regression assumption, we have the homoskedasiticity assumption given by

$$Var[\epsilon|X] = E[\epsilon^2|X] = \sigma^2.$$

To test the violation of this assumption, we want to test whether expected value of u^2 is related to one or more of the explanatory variables via certain function form (e.g., linear, quadratic, etc). Therefore, we can ultimately test the relation by regressing u^2 on predictor variables.

Methodology 15.3.3 (test for heteroskedasticity). *[5, p. 280]*

- *Estimate the model using OLS and obtain the residual estimates $\hat{e}$ and fitted value $\hat{y}$.*

- *Run one of the following regression models:*
  - $\hat{\epsilon}^2 = \delta_0 + \delta_1 x_1 + ... + \delta_k x_k + v$.
  - $\hat{\epsilon}^2 = \delta_0 + \delta_1 x_1 + ... + \delta_k x_k + \delta_{k+1} x_1^2 + \delta_{k+2} x_1 x_2 + ... + \delta_{k^2} x_k^2 + v$.
  - $\hat{\epsilon} = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + v$.
- *Use F test [Theorem 15.2.1] to see if $\delta_i, i = 1, 2, ...$ are significantly different from zero.*

15.3.3.2 *Heteroskedasticity robust estimator*

Remark 15.3.3 (motivation and general remarks).

- For a linear regression model with **known** structural error characterized by $Var[\epsilon] = \sigma^2\Sigma$, we can use generalized least square method to estimate the coefficient and its standard error [Theorem 15.1.12].
- If Ω is unknown, an
  Robust standard errors is a technique to obtain unbiased standard errors of OLS coefficients under heteroscedasticity.
- If Ω is known, another way is to use feasible GLS or feasible weighted least square.

Definition 15.3.3 (heteroscedasticity consistent standard error estimator, robust standard error estimator). *A typical heteroscedasticity consistent standard error estimator for β is defined by*

$$Var[\hat{\beta}] = (X^T X)^{-1} X^T \hat{\Sigma} X (X^T X)^{-1}$$

where $\hat{\Sigma}$ is given by

$$\hat{\Sigma} = \begin{bmatrix} \hat{\epsilon}_1^2 & & & \\ & \hat{\epsilon}_2^2 & & \\ & & \ddots & \\ & & & \hat{\epsilon}_n^2 \end{bmatrix}, \hat{\epsilon}_i^2 = (y_i - \hat{\beta} x_i)^2.$$

Some variants will multiply $\hat{\Sigma}$ by $n/n-p$ as a degree-of-freedom correction.

Remark 15.3.4. For a linear regression model with structural error characterized by $Var[\epsilon] = \sigma^2\Sigma$, the OLS estimator is given by [15.1.1]

$$Var[\hat{\beta}] = \sigma^2 (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1}.$$

15.3.3.3 *Feasible weighted least square*

Remark 15.3.5 (weighted least square as a special generalized least square). Weighted least square is the special case that $\Sigma^{-1} = diag(w_1, ..., w_p)$.

Algorithm 23: EM algorithm for least square with nonconstant variance

Input: Data set consists of X, Y

1 Start with initial weight $w_i \geq 0, i = 1, ..., p$ and the error model, for example $var(\epsilon_i) = \gamma_0 + \gamma_1 x_1$.

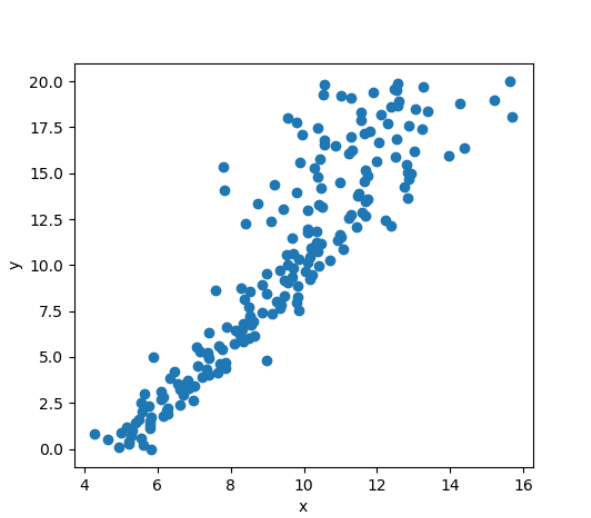
2 Use generalized least square to estimate β.

3 Use the residuals to estimate γ, by regressing x on the residual $\hat{\epsilon}^2$.

4 Re-compute the weights and go to step 2.

Output: The coefficients β

Example 15.3.2. Consider a simple regression problem for the following observations



with is generated from $y = 5 + 0.5x + w(x)Z, Z \sim N(0, 0.25)$ and $w(x) = 1$ if $0 < x < 12$ else $w(x) = 3$.

The estimated coefficients based on OLS ($\hat{\beta} = (X^T X)^{-1} X^T Y$)and WLS ($\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$) are given in the following

```
OLS result
                 coef      std err      t          P>|t|
-----------------------------------------------------------
const          5.1420      0.141     36.523      0.000
x1             0.4847      0.012     39.802      0.000
```

```
============================================================


WLS result
                coef      std err      t          P>|t|
------------------------------------------------------------
const          5.0957      0.082     62.054      0.000
x1             0.4951      0.010     47.732      0.000
============================================================
```

15.3.4 Residual normality test

15.3.4.1 *Jarque-Bera test*

Definition 15.3.4 (Jarque–Bera test). *Jarque–Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution. The test statistic JB is defined as*

$$JB = \frac{n-k-1}{6}(S^2 + \frac{1}{4}(C-3)^2),$$

where n is the number of observations, S is the sample skewness, C is the sample kurtosis, and k is the number of regressors (excluding the constant regressor). Note that for a normal distribution, $S = 0$ and $C = 3$.

Remark 15.3.6. If the data comes from a normal distribution, the JB statistic asymptotically has a $\chi^2(2)$ distribution, so the statistic can be used to test the hypothesis that the data are from a normal distribution.

15.3.4.2 *D'Agostino's K^2 test*

Definition 15.3.5 (D'Agostino's K^2 test). *D'Agostino's K^2 test is a goodness-of-fit test of whether the sample data's distribution deviate from normality. The statistic K^2 is defined by*

$$K^2 = Z_1(g_1)^2 + Z_2(g_2)^2,$$

where Z_1, Z_2 are some transformation functions and g_1, g_2 are the sample skewness and sample kurtosis.

Remark 15.3.7. If the data comes from a normal distribution, the K^2 statistic asymptotically has a $\chi^2(2)$ distribution, so the statistic can be used to test the hypothesis that the data are from a normal distribution.

15.3.5 Autocorrelation of errors

15.3.5.1 *Motivation and general remarks*

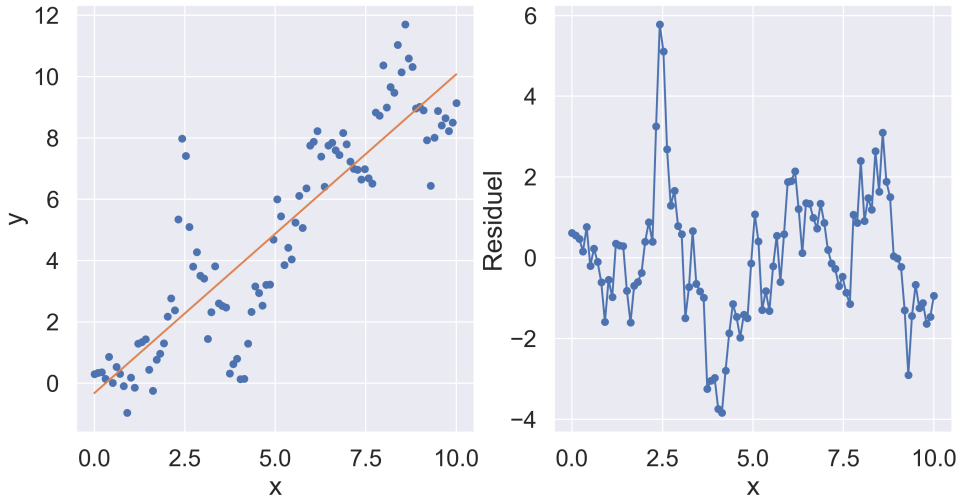
In modeling sequential observations, linear regression approach could encounter an issue known as autocorrelation of error. Specifically, we the residual resulted $\hat{\epsilon}_i = (y_i - \hat{y}_i)$ from the model is auto-correlated, rather than independent to each other. The autocorrelation of errors violate the standard assumptions[Assumption 15.1]. The consequences of auto-correlated error could be understood in the structural error framework in 15.1.1. OLS estimator is still unbiased. However, it is no longer efficient and t and F tests are no longer valid.

In Figure 15.3.2, we demonstrates linear regression results on observations containing positive and negative auto-correlated disturbance. For independent disturbance, we expect the residual distributing evenly around zero, instead of patterns (memory effect or mean-reverting) we see here.

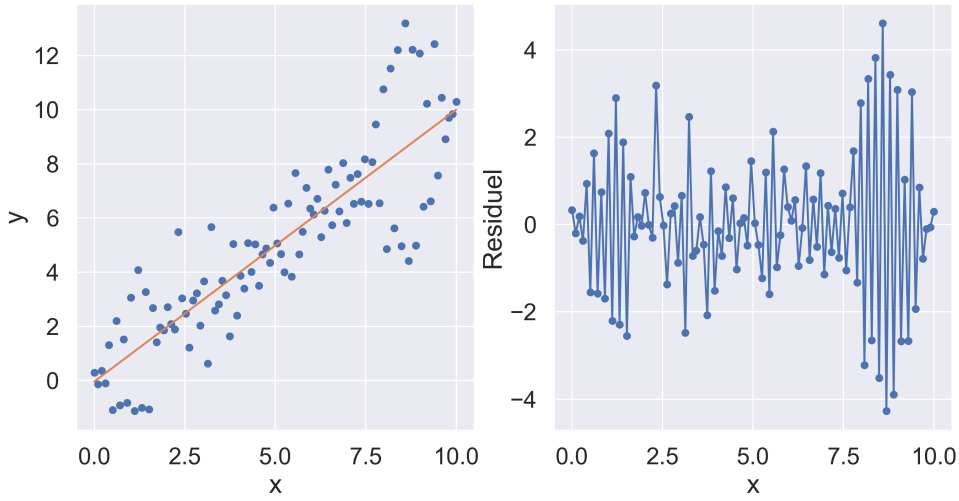
Remark 15.3.8 (source of auto-correlated error)**.** Suppose we have built a linear regression model to explain the GDP Y using capital K, labor L, and technology L over a period of time. The model is given by

$$\log(Y) = \beta_0 + \beta_1 \cdot K + \beta_2 \cdot L + \beta_3 \cdot L + \epsilon.$$

The unmodeled effects, such as policy effect, is adsorbed into the error term ϵ. However, because policy in previous years usually can have impact on subsequent years, their effects, as included in ϵ, can display autocorrelation effects.



(a) Positive auto-correlated error with correlation coefficient $\rho = 0.8$.



(b) Negative auto-correlated error with correlation coefficient $\rho = -0.8$.

Figure 15.3.2: Demonstrations on linear regression with auto-correlated error. Observations are generated by $y_i = x_i + \epsilon_i, \epsilon_i = \rho\epsilon_{i-1} + z, z \sim N(0,1)$.

15.3.5.2 *Test of autocorrelation of errors*

Beside a glance on the residual as the first diagnosis, we can also perform linear regression on the residuals and then carry out corresponding t or F tests as a more formal diagnosis procedure. Below is an example method.

Methodology 15.3.4 (linear regression coefficient test method for AR(1)). *[5, p. 417]*

- *Run the OLS regression and obtain the OLS residual* $\hat{\epsilon}_i, i = 1, 2, ..., n$.
- *Run the regression*
$$\hat{\epsilon}_i = C + \hat{\rho}\hat{\epsilon}_{i-1}, i = 2, ..., n.$$
- *Use t test to test hypothesis* $H_0 : \rho = 0; H_1 : \rho \neq 0$.

An alternative way is to perform hypothesis test on the residuals. The **Durbin-Waston test** assumes that the errors in the regression model are generated by AR(1),

$$e_t = \phi e_{t-1} + \eta, \eta \sim WN(0, \sigma^2), |\phi| < 1,$$

and proposes to examine the quantity

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}.$$

Let's first look at the method description and then interpret it.

Methodology 15.3.5 (Durbin-Watson test). *Let* $\hat{a}$

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

where $e_t = y_t - \hat{y}_t$. *The decision rule is given as*

- *If* $d < d_{L,\alpha}$, *there is statistical evidence that the error terms are positively autocorrelated.*
- *If* $d > d_{U,\alpha}$, *there is no evidence of autocorrelation.*
- *If* $d_{L,\alpha} < d < d_{U,\alpha}$, *the test is inconclusive.*

Remark 15.3.9 (interpretation).

- The Durbin Watson statistic has the following approximation:

$$\begin{aligned} d &= \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \\ &= \frac{\sum_{t=2}^{T} e_t^2}{\sum_{t=1}^{T} e_t^2} + \frac{\sum_{t=2}^{T} e_t^2}{\sum_{t=1}^{T} e_t^2} - 2\frac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^2} \\ &\approx 1 + 1 - 2\frac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^2} \approx 2(1 - \phi). \end{aligned}$$

where ϕ is the AR(1) coefficient.

- The Durbin-Watson statistic is always between 0 and 4. A value of 2 means that there is no autocorrelation in the sample. Values approaching 0 indicate positive autocorrelation and values toward 4 indicate negative autocorrelation.

Remark 15.3.10 (higher order model of errors). There are tests for higher order model of errors, such as Breusch-Pagan test.

15.3.5.3 *Models with known autocorrelation*

If we know the autoregressive dynamics of the error generation process, we can exploit this fact formulate our linear regression model.

Definition 15.3.6 (linear regression model with AR(1) error). *[6, p. 361] The linear regression model for random variable y_t depending on non-random observation x and correlated error e_t is given by*

$$y_t = \beta_0 + \beta_1 x_t + e_t$$
$$e_t = \rho e_{t-1} + v_t$$

where

- $-1 < \rho < 1$
- $E[v_t] = 0, Var[v_t] = \sigma_v^2, Cov(v_s, v_t) = 0, \forall t \neq s.$

Lemma 15.3.2 (equivalent form for auto-correlated AR(1) noise model). *[6, p. 361] The model*

$$y_t = \beta_0 + \beta_1 x_t + e_t$$
$$e_t = \rho e_{t-1} + v_t$$

has the following equivalent forms:

- $$y_t = \beta_0 + \beta_1 x_t + \sum_{i=0}^{\infty} \rho^i v_{t-i};$$
- $$(1 - \rho B) y_t = \beta_0 + \beta_1 (1 - \rho B) x_t + v_t,$$
  where B is the lag operator;
- $$y_t = \beta_0 (1 - \rho) + \beta_1 x_t + \rho y_{t-1} - \rho \beta_1 x_{t-1} + v_t,$$

or written by

$$\tilde{y}_t = \beta_0(1-\rho) + \beta_1 \tilde{x}_t + v_t,$$

where $\tilde{y}_t = y_t - \rho y_{t-1}$ *and* $\tilde{x}_t = x_t - \rho x_{t-1}$.

Proof. (1) Note that

$$\begin{aligned}(1-\rho B)e_t = v_t \implies e_t &= (1-\rho B)^{-1} v_t \\ &= \sum_{i=0}^{\infty} \rho^i v_{t-i}\end{aligned}$$

(2)(3) Multiply both sides by $(1-\rho B)$ will get the result. □

Remark 15.3.11 (alternative models). We can also model the correlated e_t using MA(q) and ARMA(p,q) model.

15.3.5.4 *Transformation to generalized linear regression*

Another modeling approach is to incorporate auto-correlated errors into the structural error model framework - generalized linear regression [Theorem 15.1.12]. Below, we first give a Lemma used to find out the structure of the error given that error has known auto-regressive properties. In the case that we do not know ρ, we can use least square to estimate the correlation [Methodology 21.2.4] first and perform iterative feasible GLS [Methodology 15.1.7]. For example, we assume the correlation parameter ρ as a solution to

$$\min_{\rho} \sum_{i=2}^{N} (e_i - \rho e_{i-1})^2$$

is given by

$$\hat{\rho} = \frac{\sum_{t=1}^{N-1} e_t e_{t+1}}{\sum_{t=1}^{N-1} e_t^2},$$

where e_i is the residue from the standard least square solution.

Lemma 15.3.3 (autoregressive transformation to weighted least square). *[2, p. 254]*

If the error term satisfying

$$\epsilon_i = \rho \epsilon_{i-1} + \xi_i, i = 2, 3, ..., n$$

where n *is the number of samples,* $|\rho| \leq 1$ *is* ***known*** *and* $E[\xi_i] = 0, E[\xi_i, \xi_j] = \sigma_0^2 \delta_{ij}$. *Then*

- $$E[\epsilon_i] = 0$$

- $$E[\epsilon_i \epsilon_j] = \sigma_0^2 \frac{\rho^{|i-j|}}{1-\rho^2}.$$

Or equivalently, the $Cov[\epsilon, \epsilon] = \sigma_0^2 V$, *where*

$$V = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & & \rho^{n-3} \\ \vdots & & & \ddots & \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}$$

Proof. Use the basic statistical properties of AR(1) processes in Lemma 21.2.5. □

Remark 15.3.12 (property of V matrix). Here we list some properties of V, which will be useful when we do weighted least square.

- $$V^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{pmatrix}$$
- $$V^{-1} = P^T P, P = \frac{1}{\sqrt{1-\rho^2}} \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{pmatrix}$$

15.3.6 Outliers analysis and robust linear regression

15.3.6.1 *Outliers and influential points*

An **outlier** is usually a data point that is distant from rest of the data. Outliers usually break a Pattern or trend formed by other data point. Informally, a data point has **high leverage** if it has "extreme" predictor x values.A data point is **influential** if it affects the predicted responses significantly via the estimated slope coefficients Outliers and high leverage data points have the potential to be influential, but not necessarily so. Figure 15.3.3 shows different types of outliers.

Using OLS results, we have more formally ways to quantify the **leverage** concept.

Definition 15.3.7 (leverage). *Let H be the orthogonal projector matrix in the multiple linear regression. H_{ii} is called the **leverage** because H_{ii} quantifies the influence that the observed response y_i has on the predicted response $\hat{y}_i$ since*

$$\hat{y}_i = H_{i1}y_1 + \cdots + H_{ii}y_i + \cdots + H_{in}y_n.$$

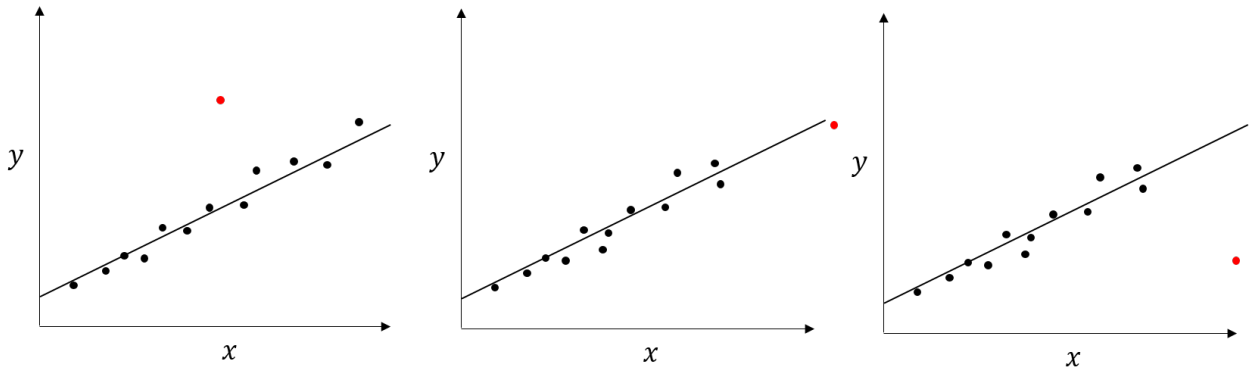
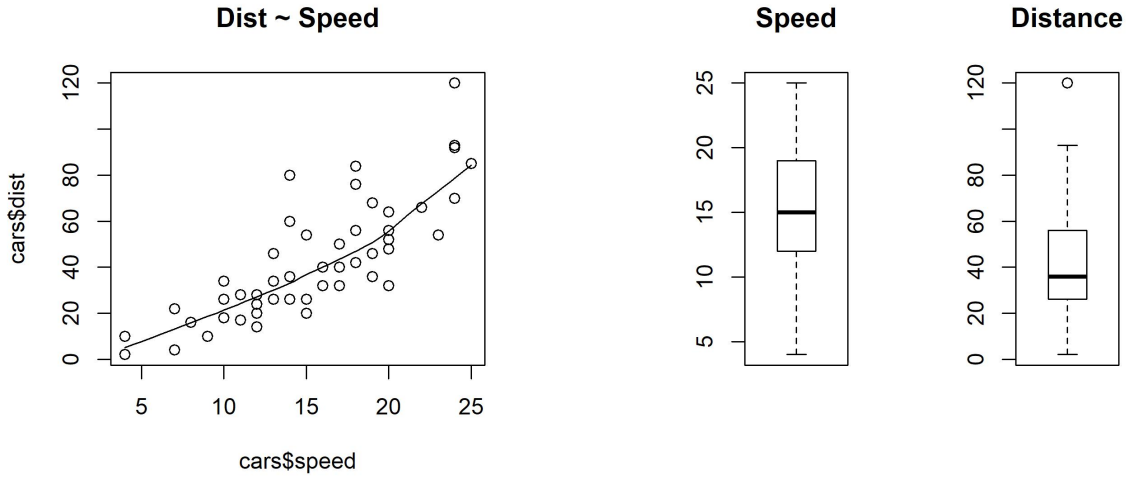


Figure 15.3.3: Illustration of an outlier, a high-leverage point, and a influential point. Left subfigure shows a red-colored outlier, which does not have high leverage and large influence on the regression result. Middle subfigure shows a red-colored high-leverage point, which is not an outlier or influential point due to its weak influence on the regression result. Right subfigure shows an influential point that is both an outlier and a high-leverage point.

Note 15.3.1 (interpretation and usage of leverage to identify influential points).

- That is, if H_{ii} is small, then the observed response y_i plays only a small role in the value of the predicted response $\hat{y}_i$; On the other hand, if H_{ii} is large, then the observed response y_i plays a large role in the value of the predicted response $\hat{y}_i$.
- H_{ii} is between 0 and 1, inclusively; and $\sum_i H_{ii} = p$. [Lemma 15.1.1]
- If
$$H_{ii} > 3\frac{\sum_i H_{ii}}{n} = 3\frac{p}{n},$$
that is, H_{ii} is more than 3 times larger than the mean leverage value, we identify the pair (x_i, y_i) as the influential points.
- Note that high influential points are not necessarily outliers.

Initial screening of outliers can be achieved by scatter plot and box plot, as showed Figure 15.3.4. The distance variable in the car R dataset has one outlier.



(a) Scatter plot. **(b)** Box plot.

Figure 15.3.4: Visual scatter and box plots to identify outliers.

15.3.6.2 *Outlier impact analysis*

As we discussed in Figure 15.3.3, an outlier might not be highly influential points. In practice, we are primarily concerned with outliers that exert significant impacts on our fitting results. One way to quantify the influence of a potential outlier is **leave-one-out** analysis, where we compare the fitting results before and after we remove one potential outlier.

To start with, we first examine the impact of data point removal on fitting results.

Lemma 15.3.4 (leave-one-out analysis on fitted coefficients). *[7, p. 268] Let $\hat{\beta}$ and $\hat{\beta}^{(i)}$ denote the least square estimate of β with and without the i data sample. Then*

$$\hat{\beta} - \hat{\beta}^{(i)} = \frac{(X^T X)^{-1} x_i e_i}{1 - h_i}$$

where e_i is the residual $y_i - \hat{\beta} x_i$, X is the design matrix, $h_i = H_{ii} = [X(X^T X)^{-1} X^T]_{ii}$, x_i is the vector representation of ith example.

Proof. Let $X(i)$ denote the design matrix X with ith row removed. Note that $X(i)^T X(i) = X^T X - x_i x_i^T$, we have

$$\begin{aligned}
[X(i)^T X(i)]^{-1} &= [X^T X - x_i x^T]^{-1} \\
&= [X^T X]^{-1} + \frac{[X^T X]^{-1} x_i x_i^T [X^T X]^{-1}}{1 - x_i^T [X^T X]^{-1} x_i} \\
&= [X^T X]^{-1} + \frac{[X^T X]^{-1} x_i x_i^T [X^T X]^{-1}}{1 - H_{ii}}
\end{aligned}$$

where we use the matrix inversion formula [] and the fact the

$$x_i^T [X^T X]^{-1} x_i = \delta_i^T X^T [X^T X]^{-1} X \delta_i = H_{ii},$$

where δ_i is unit vector with entry i being 1.

Further note that $X(i)^T Y(i) = X^T Y - x_i y_i$, we have

$$\begin{aligned}
\hat{\beta}^{(i)} =& [X(i)^T X(i)]^{-1} X(i)^T Y(i) \\
=& [X(i)^T X(i)]^{-1} (X^T Y - x_i y_i) \\
=& [X^T X]^{-1} + \frac{[X^T X]^{-1} x_i x_i^T [X^T X]^{-1}}{1 - H_{ii}} (X^T Y - x_i y_i) \\
=& \hat{\beta} - \frac{[X^T X]^{-1} x_i}{1 - H_{ii}} (y_i (1 - H_{ii}) - x_i^T \hat{\beta} + H_{ii} y_i) \\
=& \hat{\beta} - \frac{[X^T X]^{-1} x_i e_i}{1 - H_{ii}}
\end{aligned}$$

□

Now we can define a quantify known as **Cook's distance** associated with each data point. Cook's distance characterize the influence of the point on the predicted outcome results.

Definition 15.3.8 (Cook's distance). *Consider a p (including constant term) order linear regression problem. The Cook's distance for data point i is defined by*

$$D_i = \frac{(\hat{y} - \hat{y}^{(i)})^T (\hat{y} - \hat{y}^{(i)})}{p\hat{\sigma}^2}$$

where $\hat{y} = X\hat{\beta}, \hat{y}^{(i)} = X\hat{\beta}^{(i)}$.

Lemma 15.3.5 (connecting Cook's distance to residual and leverage). *The Cook's distance for data point i is given by*

$$D_i = \frac{e_i^2}{\hat{\sigma}^2} \frac{H_{ii}}{(1 - H_{ii})^2}$$

Proof. Note that

$$\hat{\beta} - \hat{\beta}^{(i)} = \frac{(X^T X)^{-1} x_i e_i}{1 - h_i},$$

and we have $(\hat{y} - \hat{y}^{(i)}) = x_i^T (\hat{\beta} - \hat{\beta}^{(i)})$ and

$$\begin{aligned} D_i &= \frac{(\hat{y} - \hat{y}^{(i)})^T (\hat{y} - \hat{y}^{(i)})}{p\hat{\sigma}^2} \\ &= \frac{(\hat{\beta} - \hat{\beta}^{(i)})^T x_i x_i^T (\hat{\beta} - \hat{\beta}^{(i)})}{p\hat{\sigma}^2} \\ &= \frac{x_i (X^T X)^{-T} x_i^T x_i (X^T X)^{-1} x_i e_i^2}{p(1 - H_{ii})^2} \\ &= \frac{H_{ii}^2 e_i^2}{p(1 - H_{ii})^2} \end{aligned}$$

□

Now we can use the metrics of leverage and Cook's distance to detect influential outliers.

Methodology 15.3.6 (practical procedures in detecting influential outlier). *In general, we can examine if a point is outlier using following methods.*

- *We can look at their leverage. A high leverage point could be a potential outlier.*
- *We can look at their studentized residuals (residual divided by estimated standard deviation), which characterizes how far they are from the regression line.*

- *We can look at their Cook's statistics, which quantify the impact on the regression result if we remove this point. We should be alter to a point whose Cook's distance is greater than 3 times of mean Cook's distance of all points.*

  Relevant software includes R package olsrr.

15.3.6.3 *Robust M-estimation linear regression*

Linear least-squares estimates can yield significantly inferior results when errors are heavy-tailed or there are outliers. One remedy is to reduce the influence of outliers via either modifying the objective function or directly discard bad observation based on some criteria. The resulting linear regression is known as robust regression.

One common robust regression is called M-estimation, which can be regarded as a generalization of maximum-likelihood estimation. In M-estimation, the estimates are determined by minimizing a particular objective function ρ that aims to reduce the influence from significantly deviating points. ρ is the function of error e, and a reasonable ρ usually satisfies

- $\rho(e) \geq 0$ and $\rho(0) = 0$.
- $\rho(e) = \rho(-e)$
- Monotonic increasing with respect to $|e|$.

With a suitable ρ function, robust regression estimator is given by

Definition 15.3.9 (robust regression M-estimator). *The robust estimator for the parameter β is obtained by solving the following optimization problem*

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(e_i) = \min_{\beta} \sum_{i=1}^{n} \rho(y_i - x_i^T \beta),$$

where the function ρ is related the likelihood function for an appropriate choice of the error distribution.

An alternative scale-invariant optimization formulation is given by

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(\frac{e_i}{s}) = \min_{\beta} \sum_{i=1}^{n} \rho(\frac{y_i - x_i^T \beta}{s}),$$

where s is a robust estimate of scale. A popular choice for s is the median absolute deviation [Definition 13.1.7]

$$s = median[e_i - median[e_i]]/0.6745.$$

Common ρ functions and their derivatives are given below [Figure 15.3.5].

Example 15.3.3 (example ρ functions and their derivatives).

- (least square)
$$\rho(e) = \frac{1}{2}e^2, \rho'(e) = e.$$
- (Huber's function)
$$\rho(e) = \begin{cases} \frac{1}{2}e^2, & |e| \le t \\ t|e| - \frac{1}{2}t^2, & |e| > t \end{cases}$$
$$\rho'(e) = \begin{cases} e, & |e| \le t \\ \frac{te}{|e|}, & |e| > t \end{cases}$$
- (Bisquare)
$$\rho(e) = \begin{cases} \frac{t^2}{6}\left(1 - \left(1 - \left(\frac{e}{t}\right)^2\right)^3\right), & |e| \le t \\ \frac{t^2}{6}, & |e| > t \end{cases}$$
$$\rho'(e) = \begin{cases} e\left(1 - \left(\frac{e}{t}\right)^2\right)^2, & |e| \le t \\ 0, & |e| > t \end{cases}$$

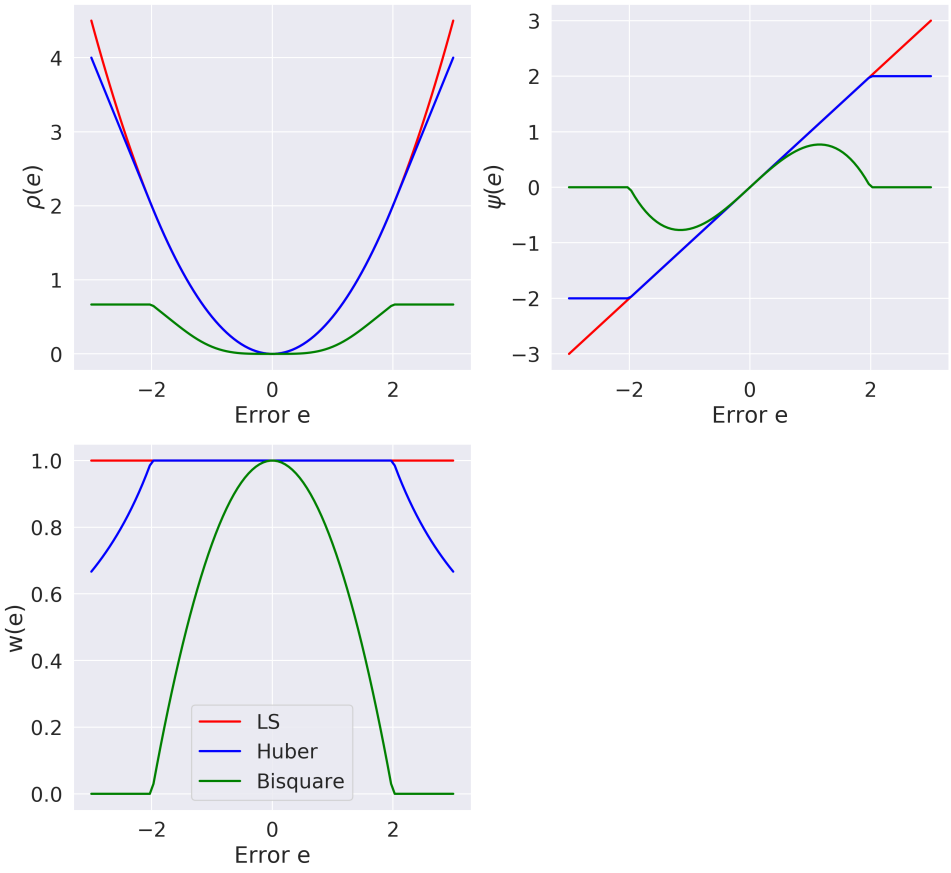


Figure 15.3.5: Different function choice for M-estimation linear regression

Remark 15.3.13 (choice of function parameters).

- In Huber and bisquare functions, smaller t gives more resistance to outliers but at the risk of discarding useful information in some samples. Particularly, the estimate for β can be of lower efficiency if the errors are truly normal.
- A guiding rule is $t = 1.345\sigma$ for the Huber and $t = 4.685\sigma$ for the bisquare, where σ is the standard deviation of the errors. Such choices produce 95-percent efficiency when the errors are normal, and still offer protection against outliers [ref].

Remark 15.3.14 (reduce to least absolute deviation (LAD) regression). If we set $t \to 0$ in Huber function, we recover the least absolute deviation (LAD) regression, which is another common tool in robust regression.

M-estimation linear regression problems are usually solved iteratively via gradient descent. The algorithm is known as **iteratively reweighed least square**, given below.

Methodology 15.3.7 (iteratively reweighed least square approach for parameter estimation). *[3, p. 373] Consider a robust estimator by solving the following problem given by*

$$\min_{\beta\in\mathbb{R}^p}\sum_{i=1}^{n}\rho(\frac{e_i}{s})=\min_{\beta}\sum_{i=1}^{n}\rho(\frac{y_i-x_i^T\beta}{s}).$$

- *The first order necessary condition for the minimum is given by*

$$\sum_{i=1}^{n}x_{ij}\psi((y_i-x_i^T\beta)/s),$$

  where $\psi=\rho'$ and x_{ij} is the i observation on the j regressor and $x_{i0}=1$.
- *(iterative reweighted algorithm) The iterative reweighted algorithm is formulated using the old estimated $\hat{\beta}_0$ and solve for the new iterate β via*

$$\sum_{i=1}^{n}x_{ij}w_{i0}\cdot(y_i-x_i^T\beta)=0, j=0,1,...,k;(*)$$

  where

$$w_{i0}=\begin{cases}\frac{\psi[(y_i-x_i^T\hat{\beta}_0)/s]}{(y_i-x_i^T\hat{\beta}_0)/s}, if\ y_i\neq x_i^T\hat{\beta}_0\\ 1\end{cases}$$

  In matrix form, $(*)$ is given by

$$X^TW_0X\beta=X^TW_0y,$$

  where W_0 is an $n\times n$ diagonal matrix of weights with diagonal elements $w_{10},w_{20},...,w_{n0}$. And the new minimizer iterate is given by

$$\hat{\beta}_1=(X^TW_0X)^{-1}X^TW_0y.$$

15.3.7 Visual diagnosis

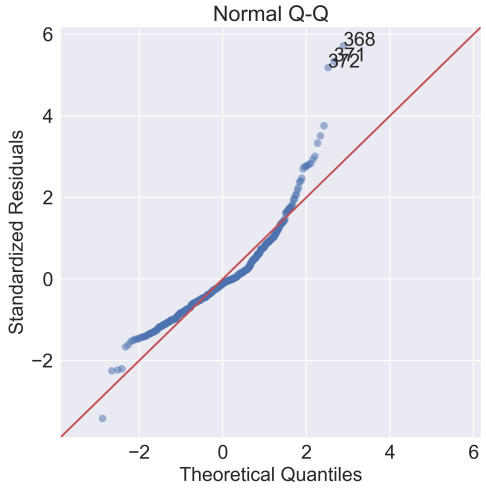
In practice, we usually use visual plots to diagnose linear regression, as showed in Figure 15.3.6. The most commonly used plots and their interpretations are the following:

- **Residuals vs Fitted plot**: A good sign of linear relation holds between predictor variables and the response variable if residuals spread relatively evenly along the

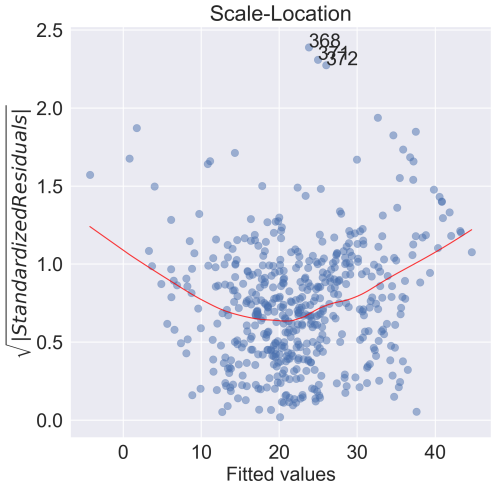
horizontal axis and the mean of residuals (the red line) is close to the horizontal line. Otherwise, there exists nonlinear relationship between the predictor variables and the response variable.

- **Normal QQ plot**: If the residuals are approximately normal, then the scatter points should be close to a straight line.
- **Scale-Location plot**: The Scale-Location plot checks if residuals are spread relatively evenly cross the range of predictors according to the assumption of equal variance (homoscedasticity).
- **Residuals vs Leverage plot**:In the Residuals vs Leverage plot, we need to pay attention to those lying around the upper right and lower right corners, which give relatively large Cook's distance (recall that Cook's distance sort of equals the product of residual and leverage). The regression results will be impacted significantly by these points Note that some points might have large residual but small leverage and some points might have large leverage but small residuals.

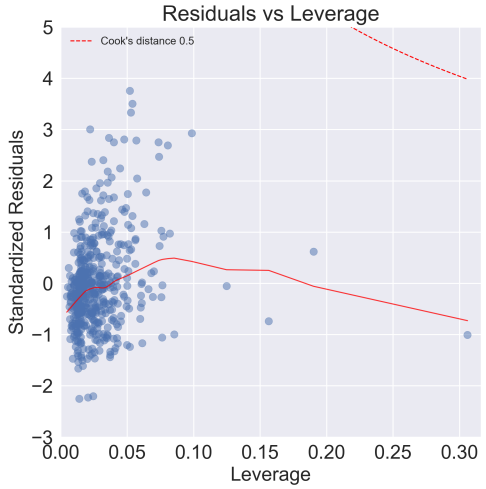
(a) Residuals vs Fitted plot.



(b) Normal QQ plot.



(c) Scale-Location plot.



(d) Residuals vs Leverage plot.

Figure 15.3.6: Common linear regression diagnosis plots

15.4 Linear regression case studies

15.4.1 Standard linear regression

We first consider a toy example by fitting a linear regression model on data generated by a model of

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \epsilon \sim N(0,1).$$

A typical fitting result summary from software like R or Python is given below.

Dep. Variable:	y	**R-squared:**	0.997
Model:	OLS	**Adj. R-squared:**	0.997
Method:	Least Squares	**F-statistic:**	1.807e+04
Date:	Sat, 01 Feb 2020	**Prob (F-statistic):**	1.73e-125
Time:	00:10:58	**Log-Likelihood:**	-125.46
No. Observations:	100	**AIC:**	256.9
Df Residuals:	97	**BIC:**	264.7
Df Model:	2		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.7795	0.468	1.666	0.099	-0.149	1.708
x1	3.0398	0.138	22.067	0.000	2.766	3.313
x2	9.9876	0.527	18.945	0.000	8.941	11.034

Omnibus:	2.287	**Durbin-Watson:**	2.070
Prob(Omnibus):	0.319	**Jarque-Bera (JB):**	1.925
Skew:	-0.338	**Prob(JB):**	0.382
Kurtosis:	3.070	**Cond. No.**	51.1

This table has the following information:

Basic information about the model fit:

- Dep. Variable: the response in the model.
- Model: the model used in the fit (OLS).
- Method: the model parameter estimation method.
- No. Observations: the sample size.
- DF Residuals: degrees of freedom of the residuals, which is sample size - number of parameters.
- DF Model: number of parameters in the model, excluding the constant term.

The goodness of fit metric

- R-squared: the coefficient of determination or the multiple correlation coefficient. [see subsubsection 15.1.7.1]
- Adj. R-squared: the adjusted coefficient of determination. [See subsubsection 15.2.2.1]
- F-statistic: the statistic for the hypothesis that if all coefficients are zero, which is a measure how significant the fit is. [See Lemma 15.1.6]

- Prob (F-statistic) the probability that you would get the above statistic, given the null hypothesis that they are unrelated. Useful for p value test.
- Log-likelihood: the log of the likelihood function [See Theorem 15.1.9].
- AIC, BIC: the Akaike Information Criterion and the Bayesian Information Criterion.

Results regarding estimation of the coefficients

- coef: the estimated value of the coefficient
- std err: the basic standard error of the estimate of the coefficient.
- t: The t-statistic value for each estimated coefficient. t > 2 is common criterion for significance.
- P > |t|: P-value that the null-hypothesis that the coefficient = 0 is true. If it is less than the confidence level, often 0.05, it indicates that there is a statistically significant relationship between the term and the response.
- 95.0% Conf. Interval: the confidence interval for estimated coefficients.

Finally, there are several statistical test results on the distribution of the residuals and multi-collinearility.

- Skewness, Kurtosis: A normal distribution will have Skewness of 0 and Kurtosis of 3.
- Omnibus, Prob(Omnibus): D'Angostino's test. It provides a combined statistical test for the presence of skewness and kurtosis. [See subsubsection 15.3.4.2]
- Jarque-Bera,Prob (JB): a different test of the skewness and Kurtosis. [See subsubsection 15.3.4.1]
- Durbin-Watson: a test for the presence of autocorrelation (that the errors are not independent.) Often important in time-series analysis [See Methodology 15.3.5]
- Cond. No: a test for multi-collinearity (if in a fit with multiple parameters, the parameters are related with each other). [See subsection 15.3.1]

We can also visual diagnosis through methods we discuss in subsection 15.3.7. As showed in Figure 15.4.1, all diagnosis plots indicates the high quality of fitted model.

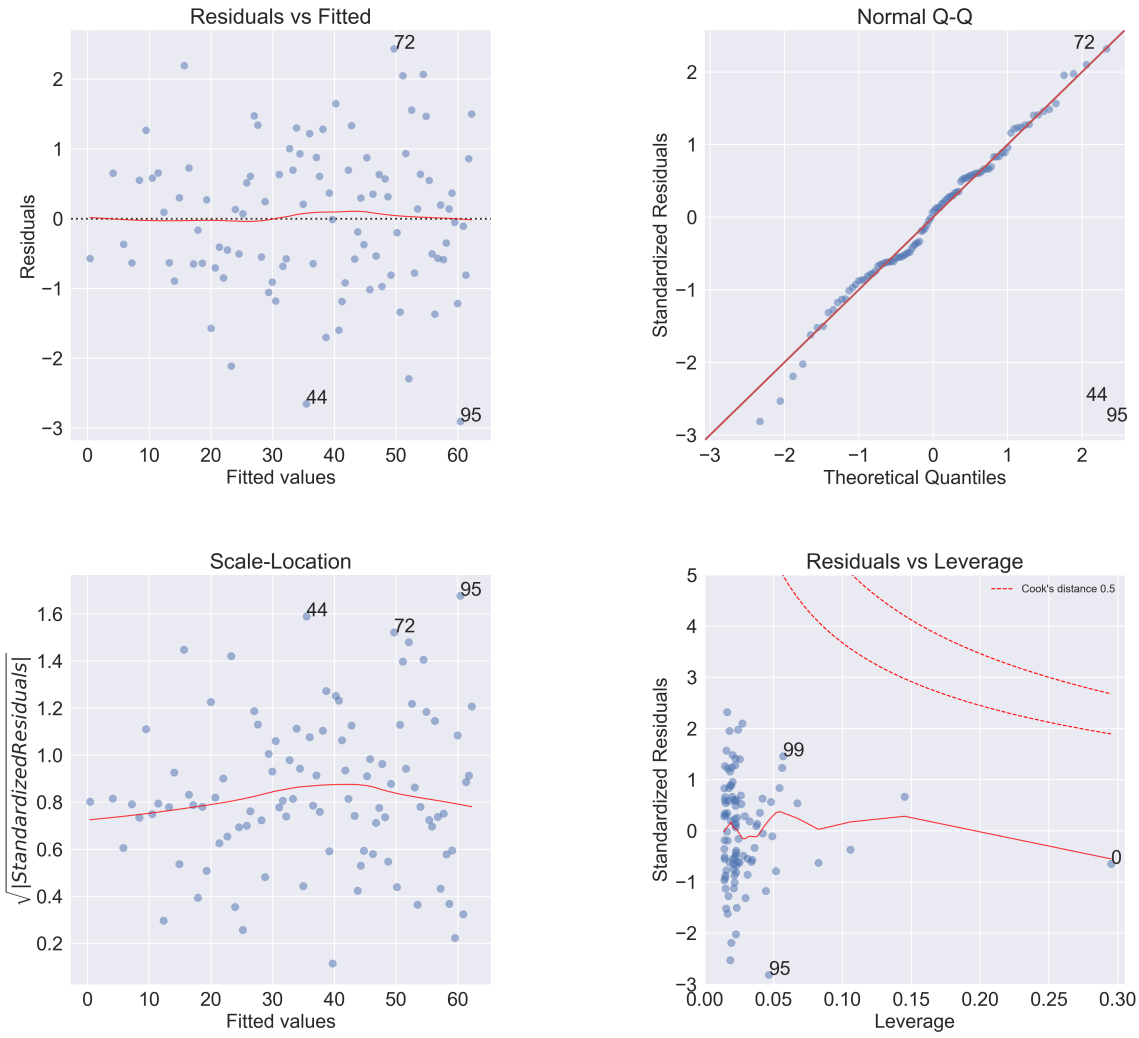


Figure 15.4.1: Diagnosis plots for a toy linear regression example

15.4.2 Boston Housing example

One classical predictive modeling problem is the Boston housing prices problem. We are given 506 samples and 13 feature variables and the goal is to model the relationship between features and the prices. The features and target variable are listed below.

- CRIM: Per capita crime rate by town
- ZN: Proportion of residential land zoned for lots over 25,000 sq. ft
- INDUS: Proportion of non-retail business acres per town.
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX: Nitric oxide concentration (parts per 10 million)

- RM: Average number of rooms per dwelling
- AGE: Proportion of owner-occupied units built prior to 1940
- DIS: Weighted distances to five Boston employment centers
- RAD: Index of accessibility to radial highways
- TAX: Full-value property tax rate per $10,000
- PTRATIO: Pupil-teacher ratio by town
- B: $1000(Bk^{-}0.63)^2$, where Bk is the proportion of [people of African American descent] by town.
- LSTAT: Percentage of lower status of the population
- **(target)** MEDV: Median value of owner-occupied homes in $1000

The linear regression model on the data set yields a R^2 of 0.741. The diagnosis plot in Figure 15.4.2 also indicates problematic issues in the linear regression model.

Dep. Variable:	0	**R-squared:**	0.741
Model:	OLS	**Adj. R-squared:**	0.734
Method:	Least Squares	**F-statistic:**	108.1
Date:	Sat, 01 Feb 2020	**Prob (F-statistic):**	6.72e-135
Time:	00:42:38	**Log-Likelihood:**	-1498.8
No. Observations:	506	**AIC:**	3026.
Df Residuals:	492	**BIC:**	3085.
Df Model:	13		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	36.4595	5.103	7.144	0.000	26.432	46.487
CRIM	-0.1080	0.033	-3.287	0.001	-0.173	-0.043
ZN	0.0464	0.014	3.382	0.001	0.019	0.073
INDUS	0.0206	0.061	0.334	0.738	-0.100	0.141
CHAS	2.6867	0.862	3.118	0.002	0.994	4.380
NOX	-17.7666	3.820	-4.651	0.000	-25.272	-10.262
RM	3.8099	0.418	9.116	0.000	2.989	4.631
AGE	0.0007	0.013	0.052	0.958	-0.025	0.027
DIS	-1.4756	0.199	-7.398	0.000	-1.867	-1.084
RAD	0.3060	0.066	4.613	0.000	0.176	0.436
TAX	-0.0123	0.004	-3.280	0.001	-0.020	-0.005
PTRATIO	-0.9527	0.131	-7.283	0.000	-1.210	-0.696
B	0.0093	0.003	3.467	0.001	0.004	0.015
LSTAT	-0.5248	0.051	-10.347	0.000	-0.624	-0.425

Omnibus:	178.041	**Durbin-Watson:**	1.078
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	783.126
Skew:	1.521	**Prob(JB):**	8.84e-171
Kurtosis:	8.281	**Cond. No.**	1.51e+04

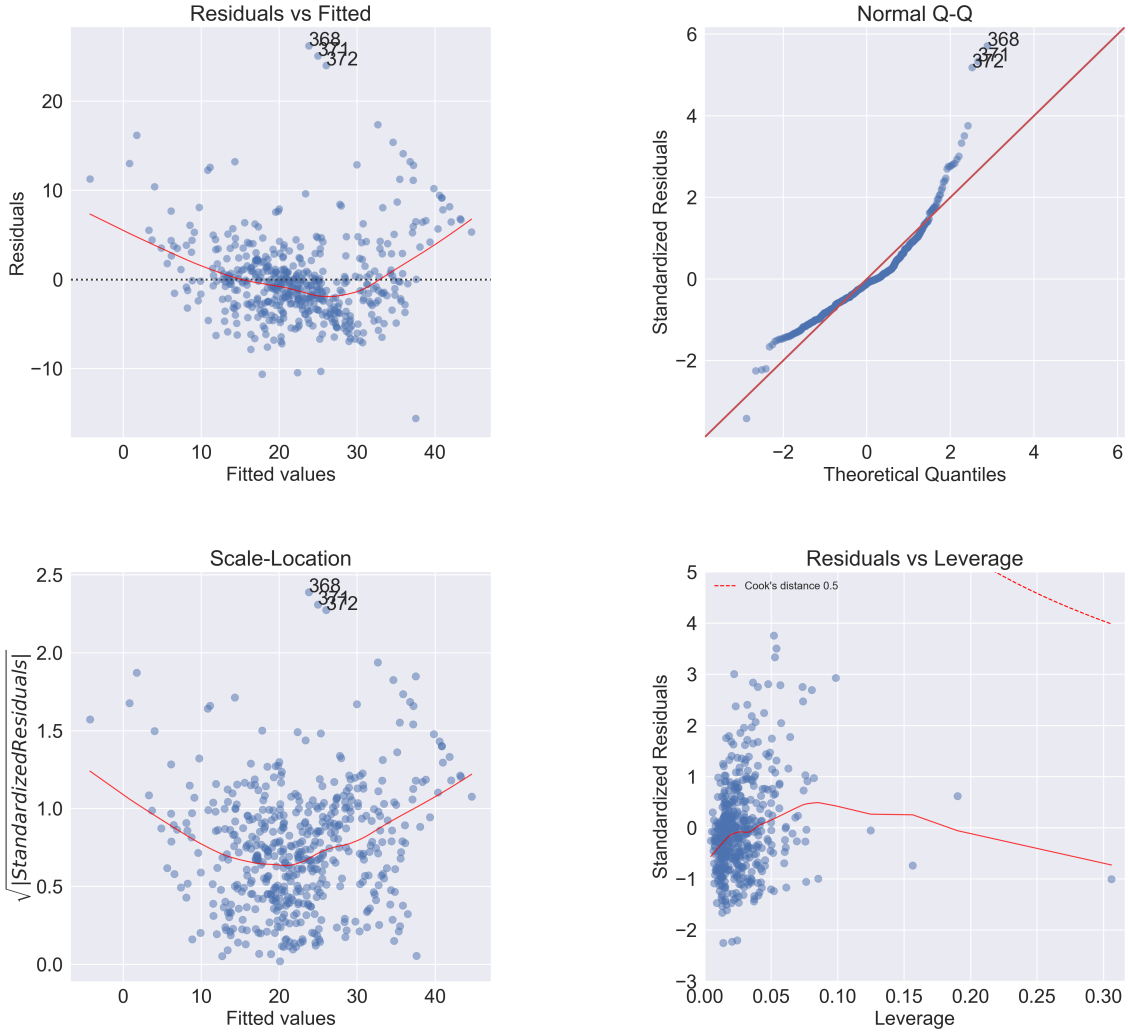


Figure 15.4.2: Diagnosis plots for the Boston Housing example

15.5 Multivariate multiple linear regression (MMLR)

15.5.1 Canonical MMLR

15.5.1.1 *Motivation and model*

In **multivariate multiple linear regression (MMLR)**, we model the linear relation between a set of outcome variables between a set of predictor variables. By contrast, in multiple linear regression, we model the linear relationship between one outcome variable and a set of predictor variables.

Example 15.5.1.

- A research is investigating the performance of cars. Speed, acceleration, and MPG are used as outcome variables to characterize performance and each car's age, price, etc. are used as predictor variables.
- A research is investigating stock market performance on a specific sector. Price returns of several key stocks are used as outcome variables and company characteristics such as cash flow, revenue generation are used as predictor variables.

Definition 15.5.1 (MMLR). *The multivariate multiple linear regression model has the form*

$$y_{ik} = \beta_{0k} + \sum_{j=1}^{p} \beta_{jk} x_{ij} + \epsilon_{ik},$$

for $i \in \{1, ..., n\}$ and $k \in \{1, ..., m\}$ where

- $y_{ik} \in \mathbb{R}$ *is the kth real-valued response for the ith observation.*
- $\beta_{0k} \in \mathbb{R}$ *is the regression intercept for kth response*
- $\beta_{jk} \in \mathbb{R}$ *is the jth predictor's regression slope for kth response*
- $x_{ij} \in \mathbb{R}$ *is the jth predictor for the ith observation*
- *random vector* $(\epsilon_{i1}, ..., \epsilon_{im}) \sim MN(0, \Sigma)$

It can be written in the following matrix form

$$Y = XB + E,$$

where

- $Y \in \mathbb{R}^{n \times m}$ *is the $n \times m$ response matrix.*
- $X = [\mathbf{1}, x_1, ..., x_p] \in \mathbb{R}^{n \times (p+1)}$ *is the $n \times (p+1)$ design matrix.*

- *B is $(p+1) \times m$ matrix of coefficients.*
- *E is the error matrix consisting of $n \times m$ random variables.*

In the expanded view, we have

$$\begin{pmatrix} y_{11} & \cdots & y_{1m} \\ y_{21} & \cdots & y_{2m} \\ y_{31} & \cdots & y_{3m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \ddots & \vdots & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_{01} & \cdots & \beta_{0m} \\ \beta_{21} & \cdots & \beta_{1m} \\ \beta_{31} & \cdots & \beta_{2m} \\ \vdots & \ddots & \vdots \\ \beta_{p1} & \cdots & \beta_{pm} \end{pmatrix} + \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ e_{21} & \cdots & e_{2m} \\ e_{31} & \cdots & e_{3m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix}$$

Remark 15.5.1 (interpret the name).

- The model is multivariate because we have $m > 1$ outcome variables.
- The model is multiple because we have $p > 1$ predictors.

Assumption 15.2 (Fundamental assumptions of the MMLR).

- *Relationship between X_j and Y_k is linear (given other predictors)*
- *x_{ij} and y_{ik} are observed random variables*
- *Disturbance $(\epsilon_{i1}, ..., \epsilon_{im}) \sim MN(0, \Sigma)$ is an unobserved random vector.*
- *βs are unknown constants.*
- *$(y_{ik}|x_{i1}, ..., x_{ip}) \sim N(\beta_{0k} + \sum_{j=1}^{p} \beta_{jk} x_{ij}, \sigma_{kk})$ for each $k \in \{1, ..., m\}$. That is, homogeneity of variance for each response.*

15.5.1.2 *Ordinary least square solution*

Theorem 15.5.1 (OLS problems).

$$\min_{B \in \mathbb{R}^{(p+1)\times m}} \|Y - XB\|_F^2 = \min_{B \in \mathbb{R}^{(p+1)\times m}} \sum_{i=1}^{n} \sum_{k=1}^{m} (y_{ik} - \beta_{0k} - \sum_{j=1}^{p} \beta_{jk} x_{ij})^2,$$

where

- *The objective function can be written by*

$$OLS(B) = \|Y - XB\|_F^2 = Tr(Y^T Y) - 2Tr(Y^T XB) + Tr(B^T X^T XB).$$

- *The first order derivative respect to B is given by*

$$\frac{\partial OLS(B)}{\partial B} = -2X^TY + 2X^TXB.$$

- *The OLS solution is given by*

$$\hat{B} = (X^TX)^{-1}X^TY;$$

The k column in $\hat{B}$ is given by

$$\hat{\beta}_k = (X^TX)^{-1}X^TY_k.$$

Proof. Note that

$$\begin{aligned} OLS(B) &= \|Y - XB\|_F^2 = Tr((Y - XB)^T(Y - XB)) \\ &= Tr(Y^TY) - Tr(Y^TXB) - Tr(B^TX^TY) + Tr(B^TX^TXB). \end{aligned}$$

Use the derivative properties for matrix trace [Lemma A.8.9], we have

$$\begin{aligned} \frac{\partial Tr(Y^TXB)}{\partial B} &= \frac{\partial Tr(B^TX^TY)}{\partial B} = X^TY \\ \frac{\partial Tr(B^TX^TXB)}{\partial B} &= 2X^TXB. \end{aligned}$$

By requiring $\frac{\partial OLS(B)}{\partial B} = 0$, we have

$$-2X^TY + 2X^TXB = 0 \implies \hat{B} = (X^TX)^{-1}X^TY.$$

□

Remark 15.5.2. For discussion in hypothesis testing, see link and [8].

15.5.2 Reduced rank regression

Sometimes we hope to the re-group a large set of original predictors into several key derived predictors, or factors, with each one expressed as the linear combination of original predictors[9][10]. Consider a centered multivariate multiple linear regression given by

$$\begin{pmatrix} y_{11} & \cdots & y_{1m} \\ y_{21} & \cdots & y_{2m} \\ y_{31} & \cdots & y_{3m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{pmatrix} = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1,K} \\ f_{21} & f_{22} & \cdots & f_{2,K} \\ f_{31} & f_{32} & \cdots & f_{3,K} \\ \vdots & \ddots & \vdots & \\ f_{n1} & f_{n2} & \cdots & f_{n,K} \end{pmatrix} \begin{pmatrix} \alpha_{01} & \cdots & \alpha_{0m} \\ \alpha_{21} & \cdots & \alpha_{1m} \\ \alpha 31 & \cdots & \alpha_{2m} \\ \vdots & \ddots & \vdots \\ \alpha_{K,1} & \cdots & \alpha_{pm} \end{pmatrix} + \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ e_{21} & \cdots & e_{2m} \\ e_{31} & \cdots & e_{3m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix}$$

or in matrix form,

$$Y = FA + E$$

where these reduced factors f_{ij} are from the linear combination of raw predictors or factors

$$f_{ij} = \sum_{k=1} x_{ik}\phi_{kj}, F = X\Phi.$$

where x_{ik} are i observation of kth predictors. So we can express the equation as

$$Y = XB + E.$$

where $B = \Phi A$, Φ is an $p \times K$ matrix and A is a $K \times m$ matrix. Then by rank of matrix product inequality [Lemma 4.4.1, also note that $K < N, K < p$], we have

$$rank(B) \leq K.$$

Such a MMLR with rank restriction is known as **reduced rank regression**, which can be used under different application scenarios:

- First, one can use it for regularization purposes like principal component linear regression [Methodology 15.3.2].
- Second, one can use it as a dimensionality reduction tool to seek most effective latent predictors in the predictor space that accounts for the variations of outcome variables.

Taken together, we have the modified optimization problem summarized in the following.

Definition 15.5.2 (reduced rank regression). *Consider a multivariate multiple linear regression. Let X and Y be centered predictor $(n \times m)$and response $(n \times q)$ data matrices.*

Then ordinary least squares (OLS) solution for reduced-rank regression can be formulated as minimizing the following cost function:

$$\min_{B\in\mathbb{R}^{(p)\times m}, rank(B)\leq r} \|Y - XB\|_F^2 = \min_{B\in\mathbb{R}^{(p+1)\times m}} \sum_{i=1}^{n}\sum_{k=1}^{m}(y_{ik} - \beta_{0k} - \sum_{j=1}^{p}\beta_{jk}x_{ij})^2.$$

Remark 15.5.3 (difference to principal component linear regression (PCLR)).

- In PCLR, we reduce the dimensionality/number of predictor variables via optimization

$$\min_{F, F^TF=I} \left\| X^TX - F\Lambda F^T \right\|_F,$$

  which enables finding low-dimensional representation of the predictor observations.
- In reduced rank regression, we combine dimensionality reduction and prediction into a single optimization framework, which usually yields better performance.

Theorem 15.5.2 (OLS solution to reduced rank regression). *[9] Consider a multivariate multiple linear regression with $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times m}$. Assume (X^TX) is non-singular. Then the minimizer for the optimization*

$$\min_{B\in\mathbb{R}^{p\times m}, rank(B)\leq r} \|Y - XB\|_F^2 = Tr[(Y - XB)^T(Y - XB)]$$

is given by

$$\hat{B} = (X^TX)^{-1}X^TYV_rV_r^T,$$

where $V_r \in \mathbb{R}^{m\times r}$ whose columns are the top r eigenvectors of $Y^TX(X^TX)^{-1}X^TY$.

Proof.

$$\begin{aligned}
&\|Y - XB\|^2 \\
=&Tr[(Y - XB)^T(Y - XB)] \\
=&Tr[(Y^TY - Y^TXB - B^TX^TY + B^TX^TXB)] \\
=&Tr[Y^TY - Y^TX(X^TX)^{-1}X^TY] + Tr[((X^TX)^{-1/2}X^TY - \\
&(X^TX)^{1/2}B)^T((X^TX)^{-1/2}X^TY - (X^TX)^{1/2}B)] \\
=&Tr[Y^TY - Y^TX(X^TX)^{-1}X^TY] + \left\|(X^TX)^{-1/2}X^TY - (X^TX)^{1/2}B\right\|_F^2
\end{aligned}$$

Based on low rank approximation results [Theorem 4.9.4], $(X^TX)^{1/2}B$ should equal $U_r\Sigma_rV_r$, which is top r SVD of $(X^TX)^{-1/2}X^TY$. Based on SVD theory [Theorem 4.9.1], V_r columns are the top r eigenvectors of

$$[(X^TX)^{-1/2}X^TY]^T[(X^TX)^{-1/2}X^TY] = Y^TX(X^TX)^{-1}X^TY.$$

On the other hand,

$$U_r \Sigma_r V_r = U \Sigma V^T V_r V_r^T,$$

where $(X^T X)^{-1/2} X^T Y = U \Sigma V^T$.

Finally,

$$(X^T X)^{1/2} \hat{B} = (X^T X)^{-1/2} X^T Y V_r V_r^T;$$

Or equivalently

$$\hat{B} = (X^T X)^{-1} X^T Y V_r V_r^T.$$

□

Corollary 15.5.2.1 (OLS solution in random variables). *Suppose the $(m+p)$ dimensional random vector $(Y^T, X^T)^T$ has mean vector o and covariance matrix with $\Sigma_{yx} = \Sigma'_{xy} = Cov(Y, X)$, and $\Sigma_{xx} = Var[X]$ non-singular. Then the minimizer for the optimization*

$$\min_{B \in \mathbb{R}^{p \times m}, rank(B) \leq r} \left\| E[Y - XB] \right\|_F^2 = Tr[E[(Y - XB)^T (Y - XB)]]$$

is given by

$$\hat{B} = \Sigma_{xx}^{-1} \Sigma_{xy} V_r V_r^T,$$

where $V_r \in \mathbb{R}^{m \times r}$ whose columns are the top r eigenvectors of $\Sigma_{yx} \Sigma_{xx} \Sigma_{xy}$

Example 15.5.2. Suppose we have n observation of returns from three stocks $y_{i,1}, y_{i,2}, y_{i3}, i = 1, ..., n$, and observation of two raw factors $x_{i,1}, x_{i,2}, i = 1, ..., n$. We aim to explain the return using following MMLR

$$\begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & y_{n3} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \end{pmatrix} + E.$$

where $E \in \mathbb{R}^{n \times 3}$ are the error terms.

Suppose we want to explain the returns by one factor, which is a linear combination of the two raw factor, we can formulate the reduced rank MMLR as

$$\begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & y_{n3} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} \begin{pmatrix} a_1 & a_2 & a_3 \end{pmatrix} + E.$$

After we use SVD method to solve the B with rank 1, we can recover Φ and A by solving equations(not unique!).

Instead of using reduced rank regression to seek the synthesized factor, we can also directly perform PCA on the observation matrix X to find the top principal component and thier loading as the factor. Then we conduct the normal MMLR.

15.6 Notes on Bibliography

For linear regression models, see [11][7]. For, linear models with *R* resources, see [12].

For multivariate statistical analysis, see [13][**anderson2009introduction**].

For copula, see [14][15][16][17].

For multivariate reduced-rank regression, see [8]

Computation software and libraries include:linearmodels, statsmodels, and scikit-learn.

BIBLIOGRAPHY

1. Greene, W. *Econometric Analysis* ISBN: 9780134461366 (Pearson, 2017).
2. Theil, H. *Principles of econometrics* (1971).
3. Montgomery, D. C., Peck, E. A. & Vining, G. G. *Introduction to linear regression analysis* (John Wiley & Sons, 2012).
4. *probability and mathematical statistic* (Zhejiang University, 2008).
5. Wooldridge, J. M. *Introductory econometrics: A modern approach* (Nelson Education, 2015).
6. Hill, R., Griffiths, W. & Lim, G. *Principles of Econometrics, 4th Edition* ISBN: 9781118136966 (John Wiley & Sons, Incorporated, 2010).
7. Seber, G. A. & Lee, A. J. *Linear regression analysis* (John Wiley & Sons, 2012).
8. Velu, R. & Reinsel, G. *Multivariate Reduced-Rank Regression: Theory and Applications* ISBN: 9781475728538 (Springer New York, 1998).
9. Velu, R. & Reinsel, G. C. *Multivariate reduced-rank regression: theory and applications* (Springer Science & Business Media, 2013).
10. Huang, D., Li, J. & Zhou, G. Shrinking factor dimension: A reduced-rank approach. *Available at SSRN 3205697* (2018).
11. Kutner, M., Nachtsheim, C. & Neter, J. *Applied Linear Regression Models* ISBN: 9780072955675 (McGraw-Hill Higher Education, 2003).
12. Faraway, J. J. *Linear models with R* (CRC press, 2014).
13. Johnson, R. & Wichern, D. *Applied Multivariate Statistical Analysis* ISBN: 9780131877153 (Pearson Prentice Hall, 2007).
14. Rüschendorf, L. *Mathematical Risk Analysis: Dependence, Risk Bounds, Optimal Allocations and Portfolios* ISBN: 9783642335907 (Springer Berlin Heidelberg, 2013).
15. Lindskog, F. *et al. Modelling dependence with copulas and applications to risk management* ().
16. McNeil, A. J., Frey, R. & Embrechts, P. *Quantitative risk management: Concepts, techniques and tools* (Princeton university press, 2015).
17. Cherubini, U., Luciano, E. & Vecchiato, W. *Copula methods in finance* (John Wiley & Sons, 2004).