Chapter 1

SIMPLE LINEAR REGRESSION

In this chapter, we consider the population regression model:

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

The mean of the distribution is $E(Y|X) = \beta_0 + \beta_1 X$ and the variance is $Var(Y|X) = Var(\beta_0 + \beta_1 X + \epsilon) = \sigma^2$.

1.1 Ordinary Least Squares (OLS) Estimation

Suppose we have n pairs of data, say $(Y_i, X_i), i = 1, ..., n$. Our goal is to find $\hat{\beta}_0$ (or b_0) and $\hat{\beta}_1$ (or b_1) such that the straight line, hence our model, "best fits" the data / observations.

Recall that the simple linear regression model stipulates that

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, ..., n.$$

The least-squares criterion upon which we can define the "best" or optimised model is given by

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \left\{ Y_i - (\beta_0 + \beta_1 X_i) \right\}^2.$$

In this case, $(b_0, b_1) = \arg\min_{(a,b)} S(a, b)$ should satisfy:

$$\begin{cases} \left.\frac{\partial S}{\partial \beta_0}\right|_{(b_0,b_1)} = -2\sum_{i=1}^{n}\{Y_i - (b_0 + b_1 X_i)\} = 0 \\ \left.\frac{\partial S}{\partial \beta_1}\right|_{(b_0,b_1)} = -2\sum_{i=1}^{n}\{Y_i - (b_0 + b_1 X_i)\} X_i = 0 \end{cases}$$

$$\implies \quad \begin{cases} n b_0 + b_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \\ b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \end{cases}.$$

Equivalently, we can write

$$\begin{pmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{pmatrix}.$$

The solution to the above set of OLS equations is given by:

$$b_0 = \bar{Y} - b_1\bar{X}, \text{ where } \bar{X} = n^{-1}\sum_{i=1}^{n} X_i, \;\; \bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$$

and

$$b_1 = \frac{\sum_{i=1}^{n} Y_i X_i - \frac{(\sum_{i=1}^{n} Y_i)(\sum_{i=1}^{n} X_i)}{n}}{\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2/n}.$$

Alternatively, recall that if $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then $A^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$, given that $ad - bc \neq 0$. Hence,

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \frac{1}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} \begin{pmatrix} \sum_{i=1}^{n} X_i^2 & -\sum_{i=1}^{n} X_i \\ -\sum_{i=1}^{n} X_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{pmatrix}$$

in which case

$$b_0 = \frac{\sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i Y_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}$$

and

$$b_1 = \frac{-\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i + n\sum_{i=1}^{n} X_i Y_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} = \frac{n^{-1}\sum_{i=1}^{n} X_i Y_i - (n^{-1}\sum_{i=1}^{n} X_i)(n^{-1}\sum_{i=1}^{n} Y_i)}{n^{-1}\sum_{i=1}^{n} X_i^2 - (n^{-1}\sum_{i=1}^{n} X_i)^2}.$$

To simplify the notation as well as to gain a more intuitive interpretation of the OLS estimates, we define

$$\begin{aligned} S_{XX} &= \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2/n = \sum_{i=1}^{n} (X_i - \bar{X})^2 \\ S_{XY} &= \sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)/n = \sum_{i=1}^{n} Y_i (X_i - \bar{X}) \end{aligned}$$

upon which we can express $b_1 = S_{XY}/S_{XX}$.Given these OLS estimates, we can also define the estimated residuals as $e_i \triangleq Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$ for $i = 1, ..., n$.

1.2 Properties of the OLS Estimators

1.2.1 Estimation of $\mathbf{b} = (\beta_0, \beta_1)$

Observe that the estimate b_1 can be expressed as a linear combination of Y_i's $(i = 1, \ldots, n)$ in the following sense:

$$\underbrace{b_1 = \frac{S_{XY}}{S_{XX}} = \sum_{i=1}^{n} c_i Y_i,}_{\text{linear combination}}$$

where $c_i = (X_i - \bar{X})/S_{XX}$. It follows that

$$\begin{aligned} E(b_1) &= E(\sum_{i=1}^{n} c_i Y_i) = \sum_{i=1}^{n} c_i E(Y_i) \\ &= \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i X_i. \end{aligned}$$

Due to the fact that

$$\begin{aligned} \sum_{i=1}^{n} c_i &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})}{S_{XX}} = \frac{n\bar{X} - n\bar{X}}{S_{XX}} = 0, \\ \sum_{i=1}^{n} c_i X_i &= \frac{\sum_{i=1}^{n} X_i(X_i - \bar{X})}{S_{XX}} = \frac{\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)\bar{X}}{S_{XX}} \\ &= \frac{\sum_{i=1}^{n} X_i^2 - (n^{-1}\sum_{i=1}^{n} X_i)^2}{S_{XX}} = 1, \end{aligned}$$

one can write $E(b_1) = \beta_0(0) + \beta_1 = \beta_1$; in other words, the estimator b_1 is unbiased. Similarly, for b_0, recall that

$$E(b_0) = E(\bar{Y} - b_1\bar{X}) = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 X_i) - \bar{X}\beta_1 = \beta_0 + \beta_1 \sum_{i=1}^{n} X_i/n - \bar{X}\beta_1 = \beta_0,$$

we can claim also that b_0 is an unbiased estimator for β_0. To establish the distribution of (b_0, b_1), we also need to evaluate the variances of these estimators. In particular, the variance of b_1 is given by:

$$Var(b_1) = Var(\sum_{i=1}^{n} c_i Y_i) = \sum_{i=1}^{n} c_i^2 Var(Y_i) = \sigma^2 \sum_{i=1}^{n} c_i^2.$$

Note that

$$\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n}(\frac{X_i - \bar{X}}{S_{XX}})^2 = \frac{1}{S_{XX}^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{S_{XX}}{S_{XX}^2} = \frac{1}{S_{XX}},$$

it follows that $Var(b_1) = \sigma^2/S_{XX}$ because Y_i's are independent. Correspondingly,

$$\begin{aligned} Var(b_0) &= Var(\bar{Y} - b_1\bar{X}) = Var(\bar{Y}) + \bar{X}^2 Var(b_1) - 2\bar{X}Cov(\bar{Y}, b_1) \\ &= \sigma^2/n + \bar{X}^2\frac{\sigma^2}{S_{XX}} = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right) \end{aligned}$$

due to the fact that

$$\begin{aligned} Cov(\bar{Y}, b_1) &= Cov\left\{\sum_{i=1}^{n}\frac{1}{n}Y_i, \sum_{i=1}^{n}\frac{(X_i - \bar{X})}{S_{XX}}Y_i\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{(X_i - \bar{X})Var(Y_i)}{S_{XX}} = \frac{\sigma^2}{S_{XX}}\sum_{i=1}^{n}\frac{(X_i - \bar{X})}{n} = 0. \end{aligned}$$

1.2.2 Estimation of σ^2

Since the residuals ϵ_i's are not observable [we only manage to estimate them via e_i's based on (b_0, b_1)], to estimate σ^2, we need to rely also on e_i's.

$$\begin{aligned} SS_{Res} &= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \{Y_i - (b_0 + b_1 X_i)\}^2 \\ &= \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 - b_1 S_{XY}. \end{aligned}$$

Define $SS_T = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$, then

$$SS_{Res} = SS_T - b_1 S_{XY} = SS_T - \frac{S_{XY}^2}{S_{XX}},$$

which leads to

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = E(SS_{Res}) = (n-2)\sigma^2 = MS_{Res} \quad \text{(residual mean square)}.$$

1.3 Testing Procedures

To establish the corresponding testing procedures that are statistically sound and justified, we consider some of the following observations:

1. Z_i are independent, then $(Z_1 + Z_2 + ... + Z_k) \sim$ Normal (CLT);

2. If $Z_i \sim$ iid $N(0,1)$, then $Z_1^2 + Z_2^2 + ... + Z_k^2 \sim \chi_k^2 \qquad \forall k \in \mathbb{N}^+$ and

3. If $Z \sim N(0,1), Y \sim \chi_m^2$, then $Z/\sqrt{Y/m} \sim t_m$.

1.3.1 Inference for $\mathbf{b}$

Since $Y_i \mid X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, then b_1 is normally distributed. As we have shown

$$E(b_1) = \beta_1 \quad \text{and} \quad Var(b_1) = \frac{\sigma^2}{S_{XX}},$$

the statistic

$$z = \frac{b_1 - \beta_{10}}{\sqrt{\sigma^2/S_{XX}}} \sim N(0,1).$$

Noteworthy, however, σ^2 is unknown. Given that $MS_{Res}(n-2)/\sigma^2 \sim \chi_{n-2}^2$ distribution and ME_{Res} and b_1 are independent [see Basu Theorem or direct calculation], we obtain

$$t \triangleq \frac{b_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{XX}}} \sim t_{n-2} \quad \text{under } H_0 \quad [\text{where } \beta_1 = \beta_{10}].$$

This gives the testing procedure where H_0 should be rejected at level α if $|t| > t_{\alpha/2,n-2}$, where H_0: $\beta_1 = \beta_{10}$ versus H_1: $\beta_1 \neq \beta_{10}$. Correspondingly, the $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$\left[b_1 - t_{\alpha/2,n-2}\sqrt{MS_{Res}/S_{XX}}, b_1 + t_{\alpha/2,n-2}\sqrt{MS_{Res}/S_{XX}}\right]$$

because

$$\Pr(b_1 - t_{\alpha/2,n-2}\sqrt{MS_{Res}/S_{XX}} \leq \beta_{10} \leq b_1 + t_{\alpha/2,n-2}\sqrt{MS_{Res}/S_{XX}}) = 1-\alpha.$$

One-sided test / C.I. can also be obtained in a similar fashion. To test the hypotheses about the intercept β_0, *i.e.* $H_0 : \beta_0 = \beta_{10}$ versus $H_1 : \beta_0 \neq \beta_{10}$, we observe that

$$b_0 = \sum_{i=1}^{n}\left\{\frac{1}{n} - \bar{X}\frac{X_i - \bar{X}}{S_{XX}}\right\}Y_i \sim N\left(\beta_0, \hat{\sigma}^2(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}})\right)$$

in which case

$$\frac{b_0 - \beta_0}{\sqrt{MS_{Res}(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}})}} \sim t_{n-2}.$$

The $100(1-\alpha)\%$ confidence interval for β_0 is given by

$$\left[b_0 - t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)}, b_0 + t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)}\right]$$

1.3.2 Inference for the Mean Response

Following the above derivation, we can write

$$\widehat{E(Y|X_0)} \triangleq \hat{\mu}_{Y|X_0} = \hat{b}_0 + b_1 X_0 \sim \text{normally distributed}$$

$$\begin{aligned} Var(\hat{\mu}_{Y|X_0}) &= Var(b_0 + b_1 X_0) = Var\{\bar{Y} + b_1(X_0 - \bar{X})\} \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2(X_0 - \bar{X})^2}{S_{XX}} = \sigma^2\left\{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}}\right\}, \end{aligned}$$

with the last equality follows because $Cov(\bar{Y}, b_1) = 0$ (Why?). Thus, the sampling distribution of $\hat{\mu}_{Y|X_0}$ can be characterised via

$$\frac{\hat{\mu}_{Y|X_0} - E(Y|X_0)}{\sqrt{MS_{Res}\{n^{-1} + (X_0 - \bar{X})^2/S_{XX}\}}} \sim t_{n-2}$$

through which the testing procedure and the corresponding confidence interval construction can be established.

1.3.3 Prediction of New Observations

We define

$$\hat{Y}_0 = b_0 + b_1 X_0$$

as a *point estimate* of the new value of the response Y_0. Note that $\psi = Y_0 - \hat{Y}_0 = Y_0 - (b_0 + b_1 X_0)$ is normally distributed. Furthermore, it can be shown that

$$\begin{aligned} Var(\psi) &= Var(Y_0 - \hat{Y}_0) = Var\{(\beta_0 - b_0) + (\beta_1 - b_1)X_0 + \epsilon\} \\ &= Var(Y_0|X_0) + Var(\hat{Y}_0|X_0) \\ &= \sigma^2 + \sigma^2 \left\{ \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}} \right\} = \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}} \right\}. \end{aligned}$$

1.4 Analysis of Variance (ANOVA)

In this section, our goal is to compare $E(Y|X) = \beta_0$ versus $E(Y|X) = \beta_0 + \beta_1 X$ to see which one is better? For the first model: $E(Y|X) = \beta_0$, one can estimate $\widetilde{\beta}_0$ can be by minimising $\sum_{i=1}^n (Y_i - \widetilde{\beta}_0)^2 \Rightarrow \widetilde{\beta}_0 = \bar{Y}$; the corresponding *residual sum of squares (SSE)* is

$$SSE_1 = \sum_{i=1}^n (Y_i - \widetilde{\beta}_0)^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 = S_{YY}.$$

For the second model, we can write

$$\begin{aligned} SSE_2 &= \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y} + b_1 \bar{X} - b_1 X_i)^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - 2b_1 \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) + b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= S_{YY} - 2(\frac{S_{XY}}{S_{XX}})S_{XY} + (\frac{S_{XY}}{S_{XX}})^2 S_{XX} \\ &= S_{YY} - S_{XY}/S_{XX} \end{aligned}$$

It is not difficult to see that $SSE_1 > SSE_2$. A more sensible question is "By how much larger in SSE_1, compared SSE_2, should one claim that the second model provides a better fit?" Note that large $SSR = SSE_1 - SSE_2 = (S_{XY})^2/S_{XX}$ implies that the second model explains much more variation than the first one does. After some algebra, one can shown

$$SSR = \sum_{i=1}^n \left\{ \left(\frac{X_i - \bar{X}}{\sqrt{S_{XX}}} \right) Y_i \right\}^2.$$

Observe that, by the Central Limit Theorem (CLT), the term in the above display $\sum_{i=1}^n (\frac{X_i - \bar{X}}{\sqrt{S_{XX}}}) Y_i \sim N(0, \sigma^2)$. As a result,

$$SSR \sim \sigma^2 \{N(0,1)\}^2 = \sigma^2 \chi_1^2.$$

Recall also that $\hat{\sigma}^2 = \sum_{i=1}^n \hat{e}_i^2/(n-2) \sim \sigma^2 \chi_{n-2}^2/(n-2)$, the ratio of these two quantities, namely

$$SSR/\hat{\sigma}^2 \stackrel{d}{=} \chi_1^2 / \{\chi_{n-2}^2/(n-2)\} \stackrel{d}{=} F(1, n-2),$$

where $F(u, \nu)$ denotes F distribution with degrees of freedom u and ν, for $u, \nu > 0$. In summary, we have:

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{\text{total variation}} = \underbrace{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}_{\text{variation explained by the model}} + \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{\text{remaining variation}}$$

$$\| $$

$$\sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum_{i=1}^n (Y_i - \bar{Y})^2$$

It suffices to show that $\sum_{i=1}^n (Y_i - \hat{Y}_i)\hat{Y}_i = 0$.

Observe that

$$\begin{aligned}
\sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i) &= \sum_{i=1}^n (Y_i - \bar{Y} + b_1\bar{X} - b_1 X_i)(\bar{Y} - b_1\bar{X} + b_1 X_i) \\
&= \bar{Y}\sum_{i=1}^n (Y_i - \bar{Y}) + b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 - b_1 \bar{Y} \sum_{i=1}^n (X_i - \bar{X}) - b_1 \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \\
&= (\frac{S_{XY}}{S_{XX}})^2 S_{XX} - (\frac{S_{XY}}{S_{XX}}) S_{XY} = 0.
\end{aligned}$$

ANOVA Table

	df	SS	MS	F
Regression	1	SSR	$MSR = SSR/1$	$MSR/MSE \leftarrow F(1, n-2)$
Error	$n-2$	SSE	$MSE = SSE/(n-2)$	
Total	$n-1$	SST		

Finally, we also examine the term $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$. One can show that $0 \leq R^2 \leq 1$ (b/c $SST = SSR + SSE$)

$$R^2 = \frac{SSR}{S_{YY}} = \frac{(S_{XY})^2}{S_{XX} S_{YY}} = \underbrace{r_{XY}^2}_{\text{correlation b/w X \& Y}}.$$

Chapter 3

MULTIPLE LINEAR REGRESSION

3.1 Model and Derviations

3.1.1 Multiple Linear Regression Models

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \epsilon \quad \overset{\sim(0,\sigma^2)}{\swarrow} \\
&= \sum_{i=0}^{k} \beta_i X_i + \epsilon, \quad \text{where } X_0 \equiv 1 \\
&= (1, X_1, ..., X_k) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \epsilon
\end{aligned} \tag{1}
$$

3.1.2 Estimation of the Model Parameters

The method of least-squares (LS) can still be used to estimate the regression coefficients in (‡). Suppose that $n > k$ observations are available (why $n > k$?), let Y_i denote the i^{th} observation for the dependent variable and x_{ij} denote the i^{th} level of regressors x_j $(j = 1, ..., k)$. We assume that ϵ has $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 < \infty$, the errors are uncorrelated.

The least squares function is:

$$
S(\beta_0, ..., \beta_k) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left\{ Y_i - (\beta_0 + \sum_{j=1}^{k} X_{ij}\beta_j) \right\}^2
$$

The function S, again, has to be minimised w.r.t. $\beta_0, ..., \beta_k$. Hence, we write

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{(b_0,...,b_k)} = -2\sum_{i=1}^{n}(Y_i - b_0 - \sum_{j=1}^{k} X_{ij}b_j) = 0$$

$$\text{and} \quad \left.\frac{\partial S}{\partial \beta_j}\right|_{(b_0,...,b_k)} = -2\sum_{i=1}^{n}(Y_i - b_0 - \sum_{j=1}^{k} X_{ij}b_j)X_{ij} = 0, \quad j = 1, ..., k$$

We have a system of $k+1$ linear equations for $k+1$ unknowns.

These equations can be expressed in a neater form:

$$\begin{array}{ccccccccc}
nb_0 & + & (\sum_{i=1}^n X_{i1})b_1 & + & (\sum_{i=1}^n X_{i2})b_2 & +\ldots+ & (\sum_{i=1}^n X_{ik})b_k & = & \sum_{i=1}^n Y_i \\
(\sum_{i=1}^n X_{i1})b_0 & + & (\sum_{i=1}^n X_{i1}^2)b_1 & + & (\sum_{i=1}^n X_{i1}X_{i2})b_2 & +\ldots+ & (\sum_{i=1}^n X_{i1}X_{ik})b_k & = & \sum_{i=1}^n X_{i1}Y_i \\
\vdots & & \vdots & & \vdots & \ddots & \vdots & & \vdots \\
(\sum_{i=1}^n X_{ik})b_0 & + & (\sum_{i=1}^n X_{ik}X_{i1})b_1 & + & (\sum_{i=1}^n X_{ik}X_{i2})b_2 & +\ldots+ & (\sum_{i=1}^n X_{ik}^2)b_k & = & \sum_{i=1}^n X_{ik}Y_i
\end{array}$$

It is more convenient to deal with this problem via matrix notation. In fact, we can write $Y_i = \beta_o + \sum_{j=1}^k \beta_j X_{ij} + \epsilon_i$ $(i = 1, ..., n)$ as $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

$$\mathbf{Y} = \underset{n\times 1 \text{ vector}}{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}, \quad X = \underset{n\times(k+1) \text{ vector}}{\begin{pmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1k} \\ 1 & X_{21} & X_{22} & \ldots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \vdots & X_{nk} \end{pmatrix}}, \quad \boldsymbol{\beta} = \underset{n\times 1 \text{ vector}}{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}} \text{ and } \quad \boldsymbol{\epsilon} = \underset{n\times 1 \text{ vector}}{\begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_k \end{pmatrix}}$$

With this notation, we can simplify (2) as $X^TX\mathbf{b} = X^T\mathbf{Y}$, which yields $\mathbf{b} = (X^TX)^{-1}X^T\mathbf{Y}$, given that $(X^TX)^{-1}$ is well defined (exists).

The fitted regression model corresponding to the levels of the regressor variable $\mathbf{X}^T = (1, X_1, ..., X_k)$ is

$$\hat{Y} = X^T\mathbf{b} = b_0 + \sum_{j=1}^{k} b_j X_j,$$

and the vector of fitted values $\hat{Y}_i$ corresponding of Y_i is

$$\hat{\mathbf{Y}} = X\mathbf{b} = X(X^TX)^{-1}X^T\mathbf{Y} \triangleq H\mathbf{Y}.$$

The $n \times n$ square matrix $H = X(X^\top X)^{-1}X^\top$ is usually called the *hat matrix*, which maps the vector of observed values $\mathbf{Y}$ into a vector of fitted values $\hat{\mathbf{Y}}$. Correspondingly, we define $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ as the n residuals.

3.1.3 Properties of the Least-squares Estimates

$$\begin{aligned} E(\mathbf{b}) &= E\{(X^\top X)^{-1}X^\top \mathbf{Y}\} = E\{(X^\top X)^{-1}X^\top (X\boldsymbol{\beta}+\boldsymbol{\epsilon})\} \\ &= E\{(X^\top X)^{-1}X^\top X\boldsymbol{\beta}\} + E\{(X^\top X)^{-1}X^\top \boldsymbol{\epsilon}\} = \boldsymbol{\beta}, \end{aligned}$$

Since $E(\boldsymbol{\epsilon}) = 0$ and $(X^\top X)^{-1}(X^\top X) = I$, we can write $Cov(\mathbf{b}) = \sigma^2(X^T X)^{-1}$.

3.1.4 Estimation of σ^2

Similar to our case in simple linear regression, it can be observed that

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}\top\mathbf{e} \\ &= (\mathbf{Y} - X\mathbf{b})^\top(\mathbf{Y} - X\mathbf{b}) = \mathbf{Y}^\top\mathbf{Y} - 2\mathbf{b}^\top X^\top\mathbf{Y} + \mathbf{b}^\top X^\top X\mathbf{b} \\ &= \mathbf{Y}^\top\mathbf{Y} - \mathbf{b}^\top X^\top\mathbf{Y} \quad \text{because} \quad (X^\top X)\mathbf{b} = X^\top\mathbf{Y}. \end{aligned}$$

The residual sum of squares has $n-p\,(=n-(k+1))$ degrees of freedom. The *residual mean square* (MSR) is

$$MSE = \frac{SSE}{n-p} = \frac{\mathbf{Y}^T\mathbf{Y} - \mathbf{b}^T X^T\mathbf{Y}}{n-k-1},$$

which is, similar to the simple regression, an unbiased estimator of σ^2.

3.1.5 Hypothesis Testing in Multiple Linear Regression

Two questions:

i. What is the *overall* quality of the model?

ii. Which specific regressor(s) seem(s) important?

- Test for significance of Regression
  To test:
  $$\mathrm{H}_0 : \beta_1 = \beta_2 = \ldots = \beta_n = 0 \quad \text{v.s.} \quad \mathrm{H}_1 : \beta_j \neq 0 \text{ for at least one } j$$
  Again, an analysis of variance (ANOVA) technique / argument is used.
  Recall that the *total sum of squares* (SST) is partitioned into a *sum of squares due to regression* (SSR) and a *residual sum of squares* (SSE), i.e.
  $$SST = \underbrace{SSR}_{SSR/\sigma^2 \sim \chi^2_k} + \underbrace{SSE}_{SSE/\sigma^2 \sim \chi^2_{n-(k+1)}}$$
  Hence $$F = \frac{SSR/k}{SSE/(n-p)} = \frac{MSR}{MSE} \sim F_{k,n-p}$$

H_0 should be rejected for large value of F. (i.e. $F > F_{\alpha,k,n-p}$)

ANOVA for Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	SSR	k	MSR	MSR/MSE
Residual	SSE	$n-k-1$	MSE	
Total	SST	$n-1$		

3.1.6 Adjusted R^2 (R_a^2)

In general, R^2 ($= 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$) always increases when additional regressors are added to the model, regardless of their contribution. Therefore, it is difficult to judge whether an increment in R^2 implies an improvement.

$\swarrow$ adj for loss of d.f.

$R_a^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$ will only increase on adding a variable to the model if the addition reduces the residual mean square.

3.1.7 Extra Sum of Squares Method

This procedure can be used to investigate the contribution of a subset of the regression variables to the model. Consider the regression model with k regressors:

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

and we want to test if some subset of $r < k$ regressors contributes significantly to the model. With a partition $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$, where β_1 is $(p-r) \times 1$ and β_2 is $r \times 1$, we are interested in testing

$$H_0 : \beta_2 = 0 \quad \text{v.s.} \quad H_1 : \beta_2 \neq 0.$$

The previous model can be written as

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon},$$

where X_i represents the columns of X associated with $\boldsymbol{\beta}_i$ (i=1,2).

For the full model, $\mathbf{b} = (X^TX)^{-1}X^T\mathbf{Y}$. $SSR(\mathbf{b}) = \mathbf{b}^TX^T - n\bar{Y}^2$ (with $p = k+1$ d.f's)and $MSE = \frac{\mathbf{Y}^T\mathbf{Y} - \mathbf{b}^TX^T\mathbf{Y}}{n-p}$. For the *reduced model*, i.e. $\beta_2 = 0$, $\mathbf{Y} = X_1\mathbf{b}_1 + \boldsymbol{\epsilon}$. Correspondingly, $\mathbf{b}_1 = (X_1^TX_1)^{-1}X_1^T\mathbf{Y}$ and the residual sum of squares $SSR(\mathbf{b}_1) = \mathbf{b}_1^TX_1^T - n\bar{Y}^2$ (with

$p - r = k + 1 - r$ d.f.'s).

The regression sum of squares due to β_2 given that β_1 is already in the model is

$$SSR(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) = SSR(\boldsymbol{\beta}) - SSR(\boldsymbol{\beta}_1)$$

with $p - (p - r) = r$ d.f.'s. It is known as the *extra sum of squares* due to $\boldsymbol{\beta}_2$.

The null hypothesis can be tested by the statistic

$$F_0 = \frac{SSR(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)/r}{MSE} \begin{array}{l} \leftarrow \text{ ratio of ind.} \\ \leftarrow \quad X^2 \text{ d.f.'s} \end{array}$$

and it is rejected if $F_0 > F_{\alpha,r,n-p}$.

Since $SST = SSR + SSE, \;\; SSR = SST - SSE$, it follows that

$$\begin{aligned} SSR(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) &= SSR(\boldsymbol{\beta}) - SSR(\boldsymbol{\beta}_1) \\ &= \{SST - SSE(\boldsymbol{\beta})\} - \{SST - SSE(\boldsymbol{\beta}_1)\} \\ &= SSE(\boldsymbol{\beta}_1) - SSE(\boldsymbol{\beta}) \end{aligned}$$

3.2 Matrix Operations and their Roles in Regression

Let $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)^T$, $f(\boldsymbol{\beta}) = f((\beta_1, \beta_2, ..., \beta_k)^T)$.

Define the derivative of f w.r.t. β as $\frac{\partial f}{\partial \boldsymbol{\beta}} = [\frac{\partial f(\boldsymbol{\beta})}{\partial \beta_1}, \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_2}, ..., \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_k}]^T$

Example 1: $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top$, $f(\boldsymbol{\beta}) = (\beta_1 + \beta_2)\beta_3$

then $\frac{\partial f}{\partial \boldsymbol{\beta}} = (\beta_3, \beta_3, \beta_1 + \beta_2)^\top$.

Example 2: $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top$, $f(\boldsymbol{\beta}) = \beta_1^2\beta_2 + log(\beta_3)$

$\frac{\partial f}{\partial \boldsymbol{\beta}} =?$ [Ans: $(2\beta_1\beta_2, \beta_1^2, 1/\beta_3)^\top$]

Results:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{c}^T\mathbf{b} = \mathbf{c} \quad , \quad \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\beta}^T\mathbf{c} = \mathbf{c} \tag{1}$$

Example 3: $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$, $f(\boldsymbol{\beta}) = \boldsymbol{\beta}^T M \boldsymbol{\beta}$, where M is 3×3 .

By product rule:

$$\frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^T M \boldsymbol{\beta}) = (\boldsymbol{\beta}^T M)^T + M\boldsymbol{\beta} = (M^T + M)\boldsymbol{\beta} \tag{2}$$

Least squares estimator :

Recall $S(\boldsymbol{\beta}) = (\mathbf{Y} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\boldsymbol{\beta}) = \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}$
To seek the minimiser $\hat{\boldsymbol{\beta}}$ such that $S()$ is minimised, we consider

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 - \underbrace{2(Y^\top X)^\top}_{\text{due to (1)}} + \underbrace{\{X^\top X + (X^\top X)^\top\}}_{\text{due to (2)}}\boldsymbol{\beta} = -2X^\top Y + \underbrace{2X^\top X\boldsymbol{\beta}}_{X^\top X \text{is symmetric}}$$

Set $\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ yields $\hat{\beta} = (X^\top X)^{-1}X^\top \mathbf{Y}$

Some calculations on Matrices

Mean:

$$E(AX) = E\begin{pmatrix} \sum_{i=1}^n a_{1i}X_i \\ \vdots \\ \sum_{i=1}^n a_{pi}X_i \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n a_{1i}E(X_i) \\ \vdots \\ \sum_{i=1}^n a_{pi}E(X_i) \end{pmatrix} = AE(X)$$

Variance:

$$\begin{aligned} Var(AX) &= E[\{AX - E(AX)\}\{AX - E(AX)\}^\top] \\ &= E[\{A(X - EX)\}\{A(X - EX)\}^\top] \\ &= AE\{(X - EX)(X - EX)^\top\}A^\top \\ &= AVar(X)A^\top \end{aligned}$$

As a reult:

$$\begin{aligned} Var(\mathbf{b}) &= Var\{(X^\top X)^{-1}X^\top \mathbf{Y}\} = (X^\top X)^{-1}X^\top Var(\mathbf{Y})X(X^\top X)^{-1} \\ &= (X^\top X)^{-1}X^\top(\sigma^2 I)X(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1} \\ &= \sigma^2(X^\top X)^{-1}. \end{aligned}$$

A Matrix Operator: trace

Trace (tr) is the sum of diagonal elements of a square matrix.

If $A \underset{m\times m}{=} \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{pmatrix}$, then $tr(A) = \sum_{i=1}^m a_{ii}$

Properties: $tr(A+B) = \sum_{i=1}^{m}(a_{ii}+b_{ii}) = tr(A)+tr(B)$

$$tr(AB) = \sum_{i=1}^{m}(\underbrace{\sum_{j=1}^{m} a_{ij}b_{ji}}_{\text{diagonal}}) = \sum_{j=1}^{m}(\sum_{i=1}^{m} b_{ji}a_{ij}) = tr(BA)$$

$$tr\{E(A)\} = \sum_{i=1}^{m} E(a_{ii}) = E(\sum_{i=1}^{m} a_{ii}) = E\{tr(A)\}$$

Now, recall the multiple regression model: $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim (\mathbf{0}, \sigma^2 I)$,

$$\begin{aligned} & \text{SSE(residual sum of squares)} \\ = \; & (\mathbf{Y} - Xb)^\top(\mathbf{Y} - X\mathbf{b}) \\ = \; & \mathbf{Y}^\top\mathbf{Y} - 2\mathbf{Y}^\top X\mathbf{b} + \mathbf{b}^\top X^\top X\mathbf{b} = \mathbf{Y}^\top\mathbf{Y} - \mathbf{Y}^\top X(X^\top X)^{-1}X^\top\mathbf{Y} \\ = \; & \mathbf{Y}^\top\{I - X(X^\top X)^{-1}X^\top\}\mathbf{Y} \end{aligned}$$

Note that

$$\begin{aligned} E(\mathbf{Y}^\top A\mathbf{Y}) &= E\{tr(\mathbf{Y}^\top A\mathbf{Y})\} = E\{tr(A\mathbf{Y}\mathbf{Y}^\top)\} = tr\{AE(\mathbf{Y}\mathbf{Y}^\top)\} \\ &= tr[AE\{(X\boldsymbol{\beta}+\boldsymbol{\epsilon})(X\boldsymbol{\beta}+\boldsymbol{\epsilon})^\top\}] \\ &= tr\{A(X\boldsymbol{\beta}\boldsymbol{\beta}^\top X^\top + \sigma^2 I)\} \\ &= tr\{A(X\boldsymbol{\beta}\boldsymbol{\beta}^\top X^\top)\} + \sigma^2 tr(A) \end{aligned}$$

Put $A = I - H = I - X(X^\top X)^{-1}X^\top$, we have

$$\begin{aligned} tr(AX\boldsymbol{\beta}\boldsymbol{\beta}^\top X^\top) &= tr\{(I - X(X^\top X)^{-1}X^\top)X\boldsymbol{\beta}\boldsymbol{\beta}^\top X^\top\} \\ &= tr\{(X - X)\boldsymbol{\beta}\boldsymbol{\beta}^\top X^\top\} = 0 \end{aligned}$$

$$\begin{aligned} tr(A) &= tr(I - X(X^\top X)^{-1}X^\top) = tr(I) - tr\{X(X^\top X)^{-1}X^\top\} \\ &= tr(I) - tr\{(X^\top X)^{-1}X^\top X\} = tr(I_n) - tr(I_{k+1}) \\ &= n - (k+1) \end{aligned}$$

$$\Rightarrow E\{SSE(\mathbf{b})\} = \sigma^2\{n-(k+1)\} \Rightarrow \hat{\sigma}^2 = \frac{SSE(\mathbf{b})}{n-(k+1)}$$

Distribution of $\mathbf{b}$

$$\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} = \mathbf{b} = (X^\top X)^{-1}X^\top\mathbf{Y} = \begin{pmatrix} \sum \mu_{0i}Y_i \\ \sum \mu_{1i}Y_i \\ \vdots \\ \sum \mu_{pi}Y_i \end{pmatrix} \sim N\left(\boldsymbol{\beta}, \sigma^2(X^\top X)^{-1}\right)$$

Distribution of $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{SSE(\mathbf{b})}{n-(k+1)} = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-(k+1)} \sim \frac{\sigma^2 \overbrace{\chi^2_{n-(k+1)}}^{\text{sum of iid normal}}}{n-(k+1)}.$$

Chapter 4

MODEL ADEQUACY CHECKING

The major assumptions made for (simple) linear regression models thus far include:

i. the relation between y (the dependent variable/response) and $\mathbf{X}$ (the explanatory variables/regressors) is (approximately) linear,

ii. the error term ϵ has zero mean and constant variance σ^2 and

iii. the errors are uncorrelated and are normally distributed.

We can always fit a regression models and obtain $\mathbf{b}$ in most cases. However, we should examine the validity of these assumptions and evaluate the adequacy of the model fitted. Model inadequacies can have potentially serious consequences and lead to misleading conclusion.

It is noteworthy that standard testing procedures introduced the previous chapters, *e.g.* t, F or R^2 statistics cannot detect such deviation from the assumptions because most of them (especially the testing procedures) are developed upon the normality assumption.

4.1 Tests for Model Assumptions

4.1.1 Tests for normality

Typically, this task can be carried out via quantile-quantile plot on the fitted residuals. Conceptually, if the residuals are normally distributed with same measured constant variance, these residuals, when plotted against $N(0, 1)$ after standardising, should form a straight line on the plot.

Here is the idea: Denote $e_{(1)} \leq e_{(2)} \leq \ldots \leq e_{(n)}$ are the ordered fitted residuals, then $e_{(i)}$ will become the $i/(n+1)$th empirical quantile of the unobservable ϵ (population error). If e's are iid samples from $\epsilon \sim N(0, \sigma^2)$, then e's empirical quantiles should match the standard normal quantile $z_i/(n+1)$ i=1,...,n, which forms a straight line.

Notice that quantile-quantile plots can only provide users with the so-called "eye-ball" tests in which case no quantitative description of how far off the distribution of the observed set of residuals deviates from the hypothesised one, hence normality. There are four typically applied tests which provide us with concrete measurements, essentially the p-values, about the deviation.

i. Kolmogorov-Smirnov Test:

For n iid observations e_i, we can define the empirical CDF F_n of ϵ as:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(e_i \leq x)$$

The Kolmogorov-Smirnov (KS) statistic for a given CDF (*i.e.* the distribution you want to test against) $F(.)$ is:

$$KS_n \triangleq \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|,$$

where $\sup_x\{f(x)\}$ can roughly interpreted as the maximum value that f(x) can achieve. Essentially, the KS statistic measures the maximum deviation between the empirical CDF versus the target CDF. If this value is large (sufficiently large in statistics sense), we should reject the hypothesis that $F_n \overset{\text{d}}{=} F$.

To obtain the corresponding threshold, we consider the Kolmogorov's Distribution, which is defined as the distribution of $K \triangleq \sup_{t\in[0,1]} |B(t)|$, where B(t) is the Brownian bridge. In particular, the cumulative distribution function of K is given by $F_K(x) = \frac{\sqrt{2\pi}}{x}\sum_{i=1}^{\infty} e^{-(2i-1)^2\pi^2/(8x^2)}$.

Under $H_0 : F_n = F$, $\sqrt{n}KS_n \xrightarrow{d} \sup_{t\in[0,1]} |B(t)|$ as $n \to \infty$. Consider an α -level test, we reject H_0 if $\sqrt{n}KS_n > k_{1-\alpha}$, where $k_{1-\alpha}$ is the value such that $P(K < k_{1-\alpha}) = 1 - \alpha, \alpha \in (0, 1)$.

ii. Cramér-von Mises Test:

Another alternative is Cramér-von Mises (CvM) test. This goodness-of-fit test statistic consider an averaged square deviation between the empirical CDF and the target one in the following manner:

$$CvM_n \triangleq \int_{-\infty}^{\infty} \{F_n(x) - F(x)\}^2 dF(x)$$

Instead of checking the maximum deviation like the KS statistic does, CvM studies a more global behaviour. Cutoff values can be obtained via simulations like standard bootstrap method.

iii. Anderson-Darling Test:

A variation of this test is called the Anderson-Darling (AD) test where an additional weight is introduced to CvM when computing the mean squared deviations:

$$AD_n \triangleq \int_{-\infty}^{\infty} \{F_n(x) - F(x)\}^2 W(x) dF(x).$$

iv. Shapiro-Wilk Test:

Shapiro-Wilk test is a test for normality (see Snapiro and Wilk, 1965):

$$SW_n \triangleq \frac{\{\sum_{i=1}^n a_i e_{(i)}\}^2}{\sum_{i=1}^n (e_i - \bar{e})^2}, \quad \text{with } (a_1, \ldots, a_n) = \frac{\mathbf{m}^\top V^{-1}}{(\mathbf{m} V^{-1} V^{-1} \mathbf{m})^{1/2}},$$

where $\mathbf{m} = (m_1, \ldots, m_n)^\top$ is a vector of the expected values of the ordered statistics of iid samples from $N(0,1)$; V is the corresponding covarianve matrix, i.e. $E\{Z_{(i)}\} = m_i$, $Cov\{Z_{(i)}, Z_{(j)}\} = V_{ij}, i, j = 1, ..., n$.

4.1.2 Tests for Constancy of Error Variance

Breusch-Pagan test: If the homogeneity of variance of ϵ does not hold, the observed residuals may be related to the explanatory variables $\mathbf{X}$. To model such potential relation, we consider an auxiliary regression of the form:

$$\mathbf{e}^2 = \gamma_0 + \boldsymbol{\gamma}_1^\top \mathbf{X} + \mathbf{V}$$

If an F-test confirms that $H_0 : \boldsymbol{\gamma}_1 = \mathbf{0}$ holds, then the notion of heteroskedasticity can be rejected.

4.2 Transformation of Data as a Remedial Measure

When the response and the regressores do not relate with each other in a linear sense, data transformation can sometimes be helpful to improve the model. There is, however, no absolute answer to which transformation is needed. In most cases, domain/subject-matter knowledge can serve as useful guide for us to determine how the response should be transformed.

4.2.1 Variance-stabalising Transformations

A common source of violation of the constant variance assumption is that the response variable y follows a distribution whose variance is related to the mean.

Example 1. *If $Y \sim Poisson(\lambda)$, $E(Y) = Var(Y) = \lambda$, since the mean of Y is related to $\mathbf{X}$, the variance of Y will also depend on a linear combination of $\mathbf{X}$ (Recall that $\hat{E}(Y|X) = X\mathbf{b}$), which clearly violates the assumption.*

Example 2. *If Y a proportion (i.e. $0 \leq Y \leq 1$), then, variance of Y will have the double-bow pattern . (Recall, If $Y \sim Bernoulli(p)$, Var$(\hat{p}) = \hat{p}(1 - \hat{p})$.)*

Observe that the variance of the square root of a Poisson random variable is independent of the mean (why? See Anscombe transformation: $A : x \mapsto 2\sqrt{x + 3/8}$), we may consider regressing $\sqrt{y}$ on $\mathbf{X}$ in our first example. Other typical transformations are tabulated as follows:

Relation b/w σ^2 to $E(Y)$	Transformation
$\sigma^2 \propto E(Y)$	$\mathbf{Y} = \sqrt{Y}$ (square root; Poisson)
$\sigma^2 \propto E(Y)\{1 - E(Y)\}$	$\mathbf{Y} = \sin^{-1}(\sqrt{Y})$ (arcsin; Binomial)
$\sigma^2 \propto \{E(Y)\}^2$	$\mathbf{Y} = \log(Y)$
$\sigma^2 \propto \{E(Y)\}^4$	$\mathbf{Y} = Y^{-1}$ (reciprocal)

4.2.2 Analytical Methods for Transformation: Box-Cox Transformation

A useful class of transformation is the power transformation $\mathbf{Y} = Y^\lambda, Y \in \mathbb{R}$. When λ approaches to 0, Y^λ tends to 1. All the information in Y will degenerate. To avoid this "discontinuity" at $\lambda = 0$, one may consider $\mathbf{Y} = \lambda^{-1}(Y^\lambda - 1)$ instead. Observe that $\lim_{\lambda \downarrow 0} \lambda^{-1}(Y^\lambda - 1) = \log(Y)$ (Why?), we can introduce the celebrated Box-Cox transformation which is based on

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(Y), & \text{otherwise.} \end{cases}$$

To standardise $Y^{(\lambda)}$ so that model summary statistics can be of more comparable scale, Box and Cox (1964) proposed the following transformation:

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda \tilde{Y}^{\lambda-1}}, & \lambda \neq 0 \\ \tilde{Y} \log(Y), & \text{otherwise,} \end{cases}$$

for $\tilde{Y} \triangleq \log^{-1}(n^{-1} \sum_{i=1}^n \log Y_i)$, which is the geometric mean of Y's.

The MLE procedure is adopted for inference on λ. The maximum likelihood estimate for λ corresponds to the value of λ such that the residual sum of squares from the fitted model SSE(λ) is minimised. This can be done by grid-search, i.e. plot SSE (λ) against λ for various choices of λ 's. We, however, should not select λ by directly comparing the SSE's from the regression $Y^{(\lambda)}$ against $\mathbf{X}$ because for each $Y^{(\lambda)}$, the SSE is measured on a different scale.

Once $\hat{\lambda}$ is obtained, we can then fit the model using $Y^{(\lambda)}$ as the response. Users are advised to choose an interpretable value of λ. c.f. $\hat{\lambda} = \dfrac{1}{2}$ and $\hat{\lambda} = 0.51236$.

References:

Box, G.E.P. and Cox, D.R. (1964). "An analysis of transformations," *Journal of the Royal Statistical Society, Series B*, 26, 211-52.

Shapiro, S.S. and Wilk, M.B. (1965). "An analysis of variance test for normality," *Biometrika, 52, 591-611.*

Chapter 5

LEVERAGE AND INFLUENCE

5.1 Leverage and Influential Points

Outliers are often identified by unusually large residuals and these observations can also affect the regression model fits.

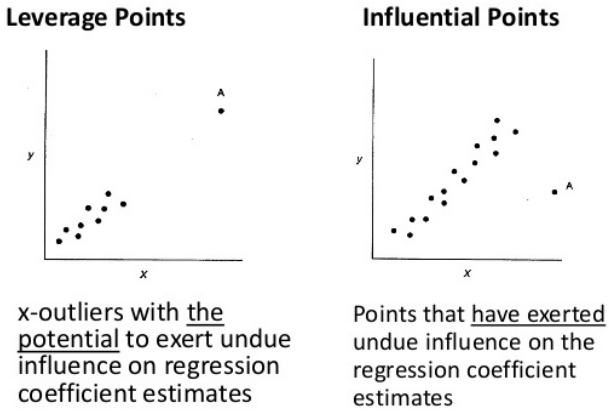


Figure 5.1: Leverage Points versus Influential Points

Figure 5.1 illustrates both a *leverage point* and an *influence point*. A leverage point has an unusual value and may control certain model properties. An influence point has a noticeable impact on the model coefficients in that it drags the regression line in its direction.

5.2 Leverage

The hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ plays an essential role in identifying influential observations. As discussed earlier in Chapter 3, H determines the variances and covariances of $\hat{\mathbf{Y}}$ and $\mathbf{e}$. (Recall that $\text{Var}(\hat{\mathbf{Y}}) = \sigma^2\mathbf{H}$, $\text{Var}(\mathbf{e}) = \sigma^2\{I - \mathbf{H}\}$). The elements h_{ij} of the matrix H may be interpreted as the amount of *leverage* exerted by the ith observation Y_i in the jth fitted values $\hat{Y}_j$. The diagonal elements

h_{ii} of H can be written as

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

where $\mathbf{X}_i$ is the i^{th} row of the design matrix X.

We can also interpret the hat matrix $\mathbf{H}$ as $\mathbf{H}$ projects $\mathbf{Y}$ onto the space spanned by $\mathbf{X}$, *i.e.* $\mathbf{HY}$ can be spanned (explained) by columns of $\mathbf{X}$ via weights $\mathbf{b}$:

$$\hat{\mathbf{Y}} = \mathbf{HY} = \mathbf{Xb} = b_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{n1} \end{pmatrix} + \ldots + b_k \begin{pmatrix} X_{1k} \\ X_{2k} \\ \vdots \\ X_{nk} \end{pmatrix}.$$

Properties of $\mathbf{H}$

i) $\mathbf{H}$ is symmetric, *i.e.* $\mathbf{H}^\top = \mathbf{H}$
because $\mathbf{H}^\top = \{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\}^\top = \mathbf{X}^\top\{(\mathbf{X}^\top\mathbf{X})^{-1}\}^\top\mathbf{X}^\top = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$

ii) $\mathbf{HH} = \mathbf{H}$ ($\mathbf{H}$ has already mapped $\mathbf{Y}$ to span($\mathbf{X}$))

because $$\begin{aligned}\mathbf{HH} &= \{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\}\{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\} = \mathbf{X}\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \\ &= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{H}\end{aligned}$$

iii) $\mathbf{HX} = \mathbf{X}$ (Project $\mathbf{X}$ on $\mathbf{X}$ should give the same $\mathbf{X}$)
because $\mathbf{HX} = \{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\}\mathbf{X} = \mathbf{X}$

iv) $(I - \mathbf{H})\mathbf{X} = 0$ (due to iii)

v) $\mathbf{H}(I - \mathbf{H}) = 0$ (due to ii)

vi) $tr(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = k + 1$
because $tr(\mathbf{H}) = tr\{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\} = tr\{(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X})\} = tr(I_{k+1}) = k + 1$

Definition: The *leverage* of i^{th} observation is the weight h_{ij} associated with Y_i in the equation:

$$\hat{Y}_i = h_{i1}Y_1 + h_{i2}Y_2 + \ldots + h_{in}Y_n$$

The leverage h_{ii} measures the influence of Y_i on its predicted value $\hat{Y}_1$. When $h_{ii} \approx 1$, $\hat{Y}_i$ is "pulled" towards Y_i.

Rule of thumb for detecting influence with leverage: the observed value of Y_i is influential if $h_{ii} > \frac{2(k+1)}{n}$. Observations with large hat diagonals and large residuals are likely to be influential.

5.3 Measures of Influence: Cook's Distance

Cook (1977, 1979) suggested a way to examine both the location of a point in X and the response variable in measuring influence. The celebrated Cook's distance is defined as

$$D_i = \frac{\{\mathbf{b}_{(i)} - \mathbf{b}\}^\top \mathbf{X}^\top \mathbf{X} \{\mathbf{b}_{(i)} - \mathbf{b}\}}{pMSE} \quad , \quad i = 1, 2, \ldots, n$$

Points with large values of D_i have substantial influence on the least-squares estimate $\mathbf{b}$. $\mathbf{b}_{(i)}$ denotes the LS estimates for the model with the i^{th} observation removed / discarded. Recall that $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$, the Cook's distance can also be expressed as:

$$D_i = \frac{\{\hat{\mathbf{Y}}_{(i)} - \mathbf{Y}\}^\top \{\hat{\mathbf{Y}}_{(i)} - \mathbf{Y}\}}{pMSE} \quad , \quad i = 1, 2, \ldots, n.$$

The above two expressions can interpret D_i as (normalised) distance between $\mathbf{b}$ and $\mathbf{b}_{(i)}$ ($\hat{\mathbf{Y}}_i$ and $\hat{\mathbf{Y}}$).

The magnitude of D_i is usually compared with $F_{\alpha,p,n-p}$. Recall that, the $(1-\alpha)$ confidence region for $\boldsymbol{\beta}$ is

$$\left\{ \boldsymbol{\beta} : \frac{(\boldsymbol{\beta} - \mathbf{b})^\top (\mathbf{X}^\top \mathbf{X})(\boldsymbol{\beta} - \mathbf{b})}{pMSE} \leq F_{\alpha,p,n-p} \right\},$$

if $D_i = F_{\alpha,p,n-p}$, it means that deleting the i^{th} observation (Y_i, X_i) will pull the estimate of $\boldsymbol{\beta}$ towards the edge of the $(1-\alpha)$ confidence region from the complete dataset $(Y_i, \mathbf{X}_i)_{i=1,\ldots,n}$

In practice, we compare D_i with $F_{0.5,p,n-p}$. Since for many F-distributions, the median (50^{th} percentile) is 1 ($q_{0.5,5,10} = 0.93$, $q_{0.5,10,10} = 1 \ldots$). Note that the distance measure D_i is not an F statistic, the cutoff of unity, however, works well in practice.

In fact, D_i can also be expressed as $D_i = \frac{(Y_i - \hat{Y}_i)^2}{pMSE} \left\{ \frac{h_{ii}}{(1-h_{ii})^2} \right\}$, $_{i=1,\ldots,n}$. With this expression, we need not run the regression model time.

To see why the expression is valid, one can write $\mathbf{b} - \mathbf{b}_{(i)} = \frac{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_i}{1 - h_{ii}} \hat{e}_i$ by observing

$$\left(\mathbf{X}^\top \mathbf{X} - \mathbf{x}\mathbf{x}^\top\right)^{-1} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} + \frac{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}\mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1}}{1 - \mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}}.$$

It follows that

$$D_i = \frac{\{\mathbf{b} - \mathbf{b}_{(i)}\}^\top \mathbf{X}^\top \mathbf{X} \{\mathbf{b} - \mathbf{b}_{(i)}\}}{p\text{MSE}} = \frac{\{(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}_i\}^\top \mathbf{X}^\top \mathbf{X} \{(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}_i\} \hat{e}_i^2}{(1-h_{ii})^2 p\text{MSE}}$$
$$= \frac{\mathbf{X}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}_i}{(1-h_{ii})^2 p\text{MSE}} \cdot \hat{e}_i^2 = \frac{h_{ii}\hat{e}_i^2}{(1-h_{ii})^2 p\text{MSE}}.$$

5.4 Measures of Influence: DFFITS and DFBETAS

Cook's distance measures is a *deletion diagnostic*. Belsley, Kuhand Welsch (1980) introduced two other measures, namely DFFITS and DFBETAS

DFFITS depicts the deletion influence of the i^{th} observation on the predicted / fitted values:

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{\text{MSE}_{(i)} h_{ii}}} \quad , \quad i = 1, 2, \ldots, n$$

Note that the denominator is just a standardisation of $\hat{Y}_i - \hat{Y}_{(i)}$ since $Var(\hat{Y}_i) = \sigma^2 h_{ii}$.

Thus, DFFITS is the member of standard deviations that the fitted value $\hat{Y}_i$ changes if the i^{th} observation is removed. Recall that $\mathbf{b} - \mathbf{b}_{(i)} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}_i e_i/(1-h_{ii})$, multiplying both sides by $\mathbf{X}_i^\top$, we obtain $\hat{Y}_i - \hat{Y}_{(i)} = \dfrac{h_{ii}e_i}{1-h_{ii}}$.

Dividing both sides of the above equation by $\sqrt{\text{MSE}_{(i)} h_{ii}}$, one can write

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{\text{MSE}_{(i)} h_{ii}}} = \frac{h_{ii}e_i}{1-h_{ii}} \left(\frac{1}{\text{MSE}_{(i)} h_{ii}}\right)^{1/2}$$
$$= \sqrt{\frac{h_{ii}}{1-h_{ii}}} e_i \sqrt{\frac{n-(k+1)-1}{\text{SEE}(1-h_{ii}) - e_i^2}}$$

$$\text{because} \quad \text{MSE}_{(i)} = \frac{1}{n-p}\sum_{j\neq i}\{Y_j - X_j^\top \mathbf{b}_{(i)}\}^2$$

$$= \frac{1}{n-p}\left[\sum_{j=1}^{n}\left\{Y_i - \mathbf{X}_j^\top \mathbf{b} + \frac{\mathbf{X}_j^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_j e_i}{1-h_{ii}}\right\}^2 - \left\{Y_i - \mathbf{X}_i^\top \mathbf{b} + \frac{h_{ii}e_i}{1-h_{ii}}\right\}^2\right]$$

$$= \frac{1}{n-p}\left\{\sum_{j=1}^{n}\left(e_j + \frac{h_{ii}e_i}{1-h_{ii}}\right)^2 - \frac{e_i^2}{(1-h_{ii})^2}\right\}$$

$$= \frac{1}{n-p}\left\{\sum_{j=1}^{n} e_j^2 + \frac{2e_i}{1-h_{ii}}\sum_{j=1}^{n} e_j h_{ij} - \frac{e_i^2}{(1-h_{ii})^2}\sum_{j=1}^{n} h_{ij}^2 - \frac{e_i^2}{(1-h_{ii})^2}\right\}$$

$$= \frac{1}{n-p}\left(\sum_{j=1}^{n} e_j^2 - \frac{e_i^2}{1-h_{ii}}\right) = \frac{(n-p)\text{MSE} - \dfrac{e_i^2}{1-h_{ii}}}{n-p-1}$$

since $H\mathbf{Y} = H\hat{\mathbf{Y}}$ ($\Rightarrow \sum_{j=1}^{n} e_j h_{ij} = 0$) and H is ($\Rightarrow \sum_{j=1}^{n} h_{ij}^2 = h_{ii}$).

Another statistic indicates how much the regression coefficient b_j's changes, in standard deviation units, if the i^{th} observation were deleted.

$$\text{DFBETAS}_{j,i} = \frac{b_j - b_{j(i)}}{\sqrt{MSE_{(i)}C_{jj}}},$$

where C_{jj} is the j_{th} diagonal element of $(\mathbf{X}^\top\mathbf{X})^{-1}$ and $b_{j(i)}$ is the j^{th} coefficient computed without the use of the i^{th} observation. If we define $\boldsymbol{R} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$, a computationally-friendly formula of DFBETAS is given by

$$\text{DFBETAS}_{j,i} = \frac{r_{j,i}}{\sqrt{\mathbf{r}_j^\top \mathbf{r}_j}} \cdot \frac{e_i}{MSE_{(i)}(1-h_{ii})},$$

if $\mathbf{r}_j^\top$ denotes the j^{th} row of $\boldsymbol{R}$ ($p \times n$ matrix).

5.5 Multicollinearity

Multicollinearity implies near-linear dependence amongst the regressors $\mathbf{X}$. Clearly, an exact linear dependence (i.e. $\mathbf{X}_i = k\mathbf{X}_j \;\; i \neq j$) would result in a singular $\mathbf{X}^\top\mathbf{X}$.

A formal method for detecting multicollinearity involves the calculation of *variance inflation factors* (VIF) for individual β parameters. One reason why the marginal t-tests on β are not significant (rejected)

is that the standard errors of the estimates s.e. (b_i) are inflated there there exists multicollinearity.

A variance inflation factor (VIF) for a β parameter is defined as

$$VIF_i = \frac{1}{1 - R_i^2} \quad , \quad i = 1, 2, \ldots, k \,,$$

where R_i^2 is the multiple coefficient of determination for $X_i = \boldsymbol{\alpha}^\top \mathbf{X}_{(i)} + \breve{\boldsymbol{\epsilon}}$.

The following indicators can be used to check for multicollinearity:

1. Significant correlations between pairs of "independent" variables in the model.
2. Non-significant t-tests for (nearly) all individual β parameters when the F test for the overall model adequacy H_0: $\beta_1 = \ldots = \beta_k = 0$ is significant (ie. inconsistent conclusion).
3. Opposite signs in the estimated parameters
4. $\text{VIF}_i > 10$

Chapter 6

LOGISTIC REGRESSION

Logistic regression models consider situations where the response variable has only two possible outcomes (dichotomous outcomes), which can be generally defined as "success" and "failure". Usually, we use 1 and 0 to denote these cases respectively. It is noteworthy that the response is essentially qualitative because the definition of success is rather arbitrary.

6.1 Models with a Binary Response Variable

Suppose that the model has the form $Y_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \epsilon_i (i = 1, \ldots, n)$, where $\mathbf{X}_i^\top = (1, X_{i1}, X_{i2}, \ldots, X_{ik})$ and $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \ldots, \beta_k)$ and the response variable Y_i takes on either 0 or 1. It is natural to assume that the response variable Y_i is a Bernoulli random variable with probability mass function shown as follows:

y_i	$P(Y_i = y_i)$
1	$\pi_i \in (0, 1)$
0	$1 - \pi_i$.

By observing that $E(\epsilon_i) = 0$, the expected value of the response variable is $E(Y_i) = 1 \times \pi_i + 0 \times (1 - \pi_i) = \pi_i$. This implies that $E(Y_i) = \mathbf{X}_i^\top \boldsymbol{\beta} = \pi_i$.

Although the above formulation is straight-forward and easy to understand, there are fundamental issues with the regression model shown above. First, note that if the response is binary, then the error terms ϵ_i can only take on two values: $\epsilon_i = 1 - \mathbf{X}_i^\top \boldsymbol{\beta}$ (when $Y_i = 1$) or $= -\mathbf{X}_i^\top \boldsymbol{\beta}$ (when $Y_i = 0$).

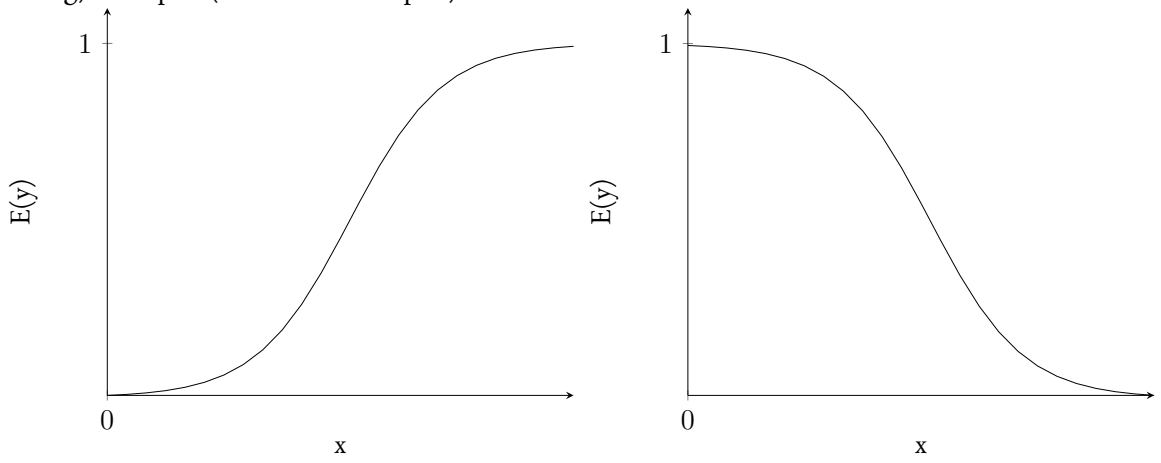
Consequently, the error terms ϵ_i cannot possibly be normal. Its variance is not constant wither because

$$\begin{aligned}\sigma_{Y_i}^2 &= E\{Y_i - E(Y_i)\}^2 = (1 - \pi_i)^2 \pi_i + (0 - \pi_i)^2 (1 - \pi_i) = \pi_i (1 - \pi_i) \\ &= E(Y_i)\{1 - E(Y_i)\}\end{aligned}$$

This indicates that the variance of the observations is a function of the mean. Finally, since $0 \leq E(Y_i) = \pi_i \leq 1$, there is a restriction on the choice of the linear response function chosen so that it maps $\mathbf{X}_i$ to

[0,1].

To overcome the aforementioned challenges, one can employ, instead, a monotonically increasing (or decreasing) S-shaped (or reverse S-shaped) function as shown below:



Examples of Logistic Response Functions

This function is regarded as the *logistic response function*, and has the form

$$\pi = E(Y|\mathbf{X}) = \frac{\exp(\mathbf{X}^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{X}^\top \boldsymbol{\beta})} \tag{1}$$

If we divide both the numerator and the denominator of the term on the right hand side of the above display, we can also write $E(y|\mathbf{X}) = \{1 + \exp(\mathbf{X}^\top \boldsymbol{\beta})\}^{-1}$. By (1), we can also see that

$$\log \frac{\pi}{1-\pi} = \log \left\{ \frac{\exp(\mathbf{X}^\top \boldsymbol{\beta})/\{1 + \exp(\mathbf{X}^\top \boldsymbol{\beta})\}}{1/\{1 + \exp(\mathbf{X}^\top \boldsymbol{\beta})\}} \right\} = \log\{\exp(\mathbf{X}^\top \boldsymbol{\beta})\} = \mathbf{X}^\top \boldsymbol{\beta}.$$

This transformation is called the *logit transformation*[1] of the probability π and the ratio $\pi/(1-\pi)$ in the transformation is called the *odds*.

6.2 Estimation

Given the general form of the logistic regression model in

$$\pi_i = E(Y_i \mid X_i) = P(Y_i = 1 \mid \mathbf{X}_i) = \frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^\top \boldsymbol{\beta})},$$

[1]There are other transformations, one of these is known as the *probit transformation*. The corresponding model is called a *probit regression model* which formulates $P(Y = 1 \mid \mathbf{X}) = \Phi(\mathbf{X}^\top \boldsymbol{\beta})$, where $\Phi(\cdot)$ denotes the cdf of a standard normal distribution.

we can use the method of *maximum likelihood* to estimate the parameters in the linear predictor $\mathbf{X}_i^\top \boldsymbol{\beta}$.

Observe that each sample observation follows the Bernoulli distribution with success probability π_i (i here specifies the dependence of the success probability on $\mathbf{X}_i$), so the p.m.f. of each observation is

$$f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \quad , \quad i = 1, \ldots, n.$$

Since the observations are independent (not i.i.d.!), the full likelihood function can be written as:

$$L(Y_1, Y_2, \ldots, Y_n \mid \boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}.$$

Maximising L is equivalent to maximising $\log L$ for $log(\cdot)$ is a monotone function. Hence, we consider the log-likelihood of $\boldsymbol{\beta}$:

$$l(\boldsymbol{\beta}) \triangleq \log L(Y_1, Y_2, \ldots, Y_n|\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right)\right\} + \sum_{i=1}^{n} \log(1-\pi_i). \qquad (\dagger)$$

Plugging in $\pi_i = \exp(\mathbf{X}_i^\top \boldsymbol{\beta})/\{1+\exp(\mathbf{X}_i^\top \boldsymbol{\beta})\}$ into (†), we yield

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i \mathbf{X}_i^\top \boldsymbol{\beta} - \sum_{i=1}^{n} \log\{1+\exp(\mathbf{X}_i^\top \boldsymbol{\beta})\}.$$

The maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{MLE}$ can be obtained by solving $\partial l(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = \mathbf{0}$. Numerical search methods could be used to compute the maximiser $\hat{\boldsymbol{\beta}}_{MLE}$ numerically.

The estimated value of the linear predictor is $\mathbf{X}_i^\top \hat{\boldsymbol{\beta}}_{MLE}$; the fitted value of the logistic regression is $\hat{Y}_i = \exp(\mathbf{X}_i^\top \hat{\boldsymbol{\beta}}_{MLE})/\{1+\exp(\mathbf{X}_i^\top \hat{\boldsymbol{\beta}}_{MLE})\}, i = 1, \ldots, n$.

6.3 Interpreting the model parameters

Consider a simplified case where there is only one regressor. The fitted value of the model at a particular value of X, say X_0, is

$$\log\left\{\frac{\hat{\pi}(X_0)}{1-\hat{\pi}(X_0)}\right\} = \hat{\beta}_0 + \hat{\beta}_1 X_0,$$

while the fitted value at $X = X_0 + 1$ is

$$\log\left\{\frac{\hat{\pi}(X_0+1)}{1-\hat{\pi}(X_0+1)}\right\} = \hat{\beta}_0 + \hat{\beta}_1 (X_0+1).$$

The difference between these two predicted values (of log odds) is $\hat{\beta}_1$. If we take antilogs, we obtain the

odds ratio:

$$\begin{aligned}\widehat{OR} &= \frac{\text{odds}(X_0+1)}{\text{odds}(X_0)} = \exp\left[\log\left\{\frac{\text{odds}(X_0+1)}{\text{odds}(X_0)}\right\}\right] \\ &= \exp[\log\{\text{odds}(X_0+1)\} - \log\{\text{odds}(X_0)\}] \\ &= \exp(\hat{\beta}_1)\end{aligned}$$

This odd ratio can be interpreted as the estimated increase in the probability of success associated with a one-unit change in the value of the regressor / predicted variable.

The interpretation of the regression coefficients in the multiple logistic regression model is similar to the case when the linear predictor contains only one regressor. The quality $\exp(\hat{\beta}_j)$ is the odds ratio for regressor X_j, assuming all other predictor variables are held fixed.

6.4 Connection with 2×2 Contingency Table

There is a close connection between the odds ratio in the logistic regression and the 2×2 contingency table (This approach is widely used for analysing categorical data).

A typical example for using the contingency table is for during testing. Consider a group of patients, who were treated with either an active drug or a placebo, ended up in one of the possible two outcomes: infected or non-infected / diseased or not diseased. A 2×2 contingency table is a neat form to present data of much format:

Response	$X_1 = 0$ (Active Drug)	$X_1 = 1$ (Placebo)
$Y = 0$ (not infected)	n_{00}	n_{01}
$Y = 1$ (infected)	n_{10}	n_{11}

Define n_{ij} ($i = 0, 1$, $j = 0, 1$) the number of patients in each cell (combination of treatments and outcomes). The odds ratio is defined as

$$\frac{\text{Proportional infected} \mid \text{active drug}}{\text{Proportion infected} \mid \text{placebo}} = \frac{n_{11}/n_{01}}{n_{10}/n_{00}} = \frac{n_{11} \cdot n_{00}}{n_{10} \cdot n_{01}}$$

Recall for the logistic model, $\log(\pi/(1-\pi)) = \beta_0 + \beta_1 X_1$. Hence, for $X_1 = 0$, we have

$$\beta_0 = \log \frac{P(Y = 1 \mid X_1 = 0)}{P(Y = 0 \mid X_1 = 0)}.$$

When $X_1 = 1$,

$$\log \frac{P(Y = 1 \mid X_1 = 1)}{P(Y = 0 \mid X_1 = 1)} = \log \frac{P(Y = 1 \mid X_1 = 0)}{P(Y = 0 \mid X_1 = 0)} + \beta_1.$$

It follows that,

$$\beta_1 = \log \frac{P(Y = 1|X_1 = 1) \cdot P(Y = 0|X_1 = 0)}{P(Y = 0|X_1 = 1) \cdot P(Y = 1|X_1 = 0)} = \log \frac{n_{11} \cdot n_{00}}{n_{01} \cdot n_{10}}.$$

The quantity $\exp(\beta_1)$ is equivalent to the odds ratio in the 2×2 contingency table.

The logistic regression model can be more general than the contingency table because it can incorporate other predictors which may affect the odds ratio. Suppose, for example, the age of each patient is recorded as well. One can then easily incorporate it to the logistic regression model: $\log(\pi/(1-\pi)) = \beta_0 + \beta_1(\text{TREATMENT}) + \beta_2(\text{AGE})$. Extension of the 2×2 contingency table is, however, not that straight-forward.

6.5 Inference on Model Parameters

Hypothesis testing in logistic regression is based on *likelihood ratio tests*. The procedure is a large sample procedure that is developed upon asymptotic theory. The key statistic is called *deviance*.

The deviance of a model compares the log-likelihood of the fitted model of interest to the log-likelihood of a saturated model, the model that has exactly n parameters that fits the sample data perfectly.

When the underlying data are Bernoulli data, $Y_i \sim Bernoulli(p_i)$, a "saturated model", *i.e.* a model in which we have a different parameter for each individual is not informative. In particular, one estimate of p_i for the saturated model is $\hat{p}_i = Y_i/1 = I(Y_i = 1)$. Recall that the likelihood of the saturated model at the MLE is $\prod_{i=1}^n \hat{p}_i^{Y_i}(1-\hat{p}_i)^{1-Y_i} = \prod_{i=1}^n 1 = 1$. Hence the corresponding log-likelihood is 0.

The deviance compares the log-likelihood of the saturated model with that of the fitted model. Specifically,

$$\text{Deviance} \quad D^2(M_1) \triangleq 2\log L(\text{saturated model}) - 2\log L(\hat{\beta}_{MLE})$$

In general, the underlying data can be viewed as made up of J binomials (which correspond to different $\mathbf{X}_i$'s values) and there are n_j subjects with the same covariate vector $\mathbf{X}_j$, $j = 1, \ldots, J$; $n \triangleq \sum_{j=1}^J n_j$. With $\mathbf{X}_j$ the covariate vector associated with group j, we want to test the fit of the model $logit(p_j) = \mathbf{X}_j^\top \boldsymbol{\beta}$, versus the saturated model which estimates a different p_j for each j individually. The likelihood ratio statistic for a given model M_1 with estimates $\hat{p}_j$ versus a "saturated" model with $\hat{p}_{j,sat} = Y_i/n_j$ is in fact the deviance:

$$\begin{aligned} D^2(M_1) &= \sum_{j=1}^J \left\{ Y_j \log\left(\frac{Y_j}{n_j\hat{p}_j}\right) + (n_j - Y_j)\log\left(\frac{n_j - Y_j}{n_j(1-\hat{p}_j)}\right)\right\} \\ &= \sum_{j=1}^J \sum_{k=1}^2 O_{jk}\log(\frac{O_{jk}}{E_{jk}}) \sim \chi^2_d \end{aligned}$$

under the null where $E_{j1} = n_j\hat{p}_j$ and $E_{j2} = n_j(1-\hat{p}_j)$ and $d = n - p$ denotes the difference between the number of parameters in the saturated model and that in M_1. If $D^2(M_1) \leq \chi^2_{\alpha,n-p}$, then

the fitted model is considered adequate; otherwise, the fitted model is concluded as inadequate.

Likelihood Hypotheses on Subsets of Parameter using Deviance

We can also use the deviance to test hypotheses or subsets of the model parameters, just as we used the partial F test statistic in the normal error linear regression model.

Since the model can be written as $\log(\frac{\pi}{1-\pi}) = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta} + \mathbf{X}_2\boldsymbol{\beta}_2$, where the full model has p parameters, $\boldsymbol{\beta}_1$ contains $p-r$ of these parameters and $\boldsymbol{\beta}_2$ contains the remaining r, we test the hypotheses: H_0: $\boldsymbol{\beta}_2 = \mathbf{0}$ vs H_1: $\boldsymbol{\beta}_2 \neq \mathbf{0}$.

We can compare the difference in deviance between these two models, say M_1 for the reduced model and M_2 the full one,

$$\Delta_D^2(M_2|M_1) = D^2(M_1) - D^2(M_2) \qquad (\ddagger)$$

and this quantity has $n-(p-r)-(n-p) = r$ degrees of freedom. If the null is true and if n is large, the difference in deviance specified in (‡) has a χ^2 distribution with r d.f.
Therefore, the decision criteria is:

$$\text{If} \quad \Delta_D^2(M_2|M_1) \leq \chi^2_{\alpha,r} \text{ , reject } H_0.$$

Sometimes $\Delta_D^2(M_2|M_1)$ is regarded as the *partial deviance*.
Finally, it is noteworthy that

$$\begin{aligned}\Delta_D^2(M_2|M_1) &= 2[\log\{L(\hat{\beta})|sat\} - \log\{L(\tilde{\beta})|M_1\}] - 2[\log\{L(\hat{\beta})|sat\} - \log\{L(\tilde{\beta})|M_2\}] \\ &= 2[\log\{L(\tilde{\beta})|M_2\} - \log\{L(\tilde{\beta})|M_1\}],\end{aligned}$$

which is also the likelihood ratio between M_1 and M_2 ($\times 2$).

Chapter 7

MODEL SELECTION

Two systematic methods are designed to reduce a large list of potential predictors to a more manageable model that is easier to interpret. These techniques are known as *variable screening procedures*, which objectively determine which independent variables are the most important predictors of Y.

7.1 Stepwise Regression

Step 1. All possible one-variable models of the form

$$E(Y \mid X_i) = \beta_0 + \beta_1 X_i$$

are fitted for $i = 1, \ldots, k$. For each model, test H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$. The independent variable that produces the largest (absolute) t-value is declared to be the best one-variable predictor of Y.

Step 2. The next step is to search through the remaining $k - 1$ independent variables for the best two-variable model of the form:

$$E(Y|\breve{X}_1, X_i) = \beta_0 + \beta_1 \breve{X}_1 + \beta_2 X_i,$$

where $\breve{X}_1$ denotes the variable chosen in the first step. Pick the variable that gives the largest t-test value. If the corresponding t-test statistic for β_1 cannot conclude that $\breve{X}_1$ is significant, $\breve{X}_1$ will be removed.

Step 3. Repeat Step 2 for the next candidate until no further independent variables can be found significant.

7.2 All-Possible-Regression Selection Procedures

Criteria:

a) R^2 Criterion

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

b) Mallow's C_p Criterion

Define the *total mean square error* (TMSE) for a fitted regression model:

$$\text{TMSE} = E\left[\sum_{i=1}^{n}\{\hat{Y}_i - E(Y_i)\}^2\right] = \overbrace{\sum_{i=1}^{n}\{E(\hat{Y}_i) - E(Y_i)\}^2}^{\text{squared bias}} + \overbrace{\sum_{i=1}^{n} Var(\hat{Y}_i)}^{\text{variance}},$$

where $E(\hat{Y}_i)$ is the mean response for the subset (fitted) regression model and $E(Y_i)$ is the mean response of the true model.

In fact, if we define $\text{SSB}_p = \sum_{i=1}^{n}\{E(\hat{Y}_i) - E(Y_i)\}^2$, where $p = k + 1$ denotes the number of parameters included in the linear model, the standardised total mean square error can be written as

$$\begin{aligned}
\Gamma_p &\triangleq \frac{\text{TMSE}}{\sigma^2} \\
&= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\{E(\hat{Y}_i) - E(Y_i)\}^2 + \sum_{i=1}^{n} Var(\hat{Y}_i)\right] \\
&= \frac{\text{SSB}_p + \sum_{i=1}^{n} Var(\hat{Y}_i)}{\sigma^2} \\
&= \frac{E(\text{SSE}_p) - (n-p)\sigma^2 + p\sigma^2}{\sigma^2} = \frac{E(\text{SSE}_p)}{\sigma^2} - n + 2p,
\end{aligned}$$

because (i) $\sum_{i=1}^{n} Var(\hat{Y}_i) = tr\{Var(\mathbf{HY})\} = \sigma^2 tr(\mathbf{HIH}) = \sigma^2 tr(\mathbf{H}) = p\sigma^2$ and (ii)

Small values of Γ_p imply that the subset regression model has a small total mean square error relative to σ^2. Since neither TMSE nor σ^2 is known, a good proxy called C_p criterion is used instead:

$$C_p = \frac{\text{SSE}_p}{\hat{\sigma}^2} + 2(k+1) - n.$$

If the p-term model has negligible bias, then $\text{SSB}_p = 0$ in which case $E(\text{SSE}_p) = (n-p)\sigma^2$, and $E(C_p \mid \text{Bias} = 0) = (n-p)\sigma^2/\sigma^2 - n + 2p = p$.

When using the C_p criterion, it can be helpful to visualise the plot of C_p as a function of p for each regression model. Regression equations with little bias will have values of C_p that fall near the linear $C_p = p$while those equations with substantial bias will fall above this line. Generally speaking, small values of C_p are desirable.

c) PRESS statistic / Predicted Residual Sum of Squares

$$\text{PRESS}_p = \sum_{i=1}^{n} \{Y_i - \hat{Y}_{(i)}\}^2 = \sum_{i=1}^{n} \left(\frac{\hat{e}_i}{1 - h_{ii}} \right)^2.$$

One selected the subset regression model based on a small value of PRESS_p.

d) Akaike Information Criterion (AIC)

Mallow's C_p is (almost) a special case of AIC, which is defined as

$$AIC(\mathcal{M}) = -2\log(\mathcal{M}) + 2p(\mathcal{M}),$$

where $\mathcal{M}$ denotes a particular linear regression model; $p(\mathcal{M})$ specifies the number of parameters in this model.

If the model is correct, then the log-likelihood of (β , σ) is

$$\log L(\beta, \sigma^2 | X, \mathbf{Y}) = -\frac{n}{2}\{\log(2\pi) + \log \sigma^2\} - \frac{1}{2\sigma^2}||\mathbf{Y} - X\boldsymbol{\beta}||^2,$$

Since $\mathbf{b} = \hat{\boldsymbol{\beta}}_{MLE}$ and $\hat{\sigma}^2_{MLE} = \frac{1}{n}\text{SSE}(\mathcal{M})$, AIC can be written as

$$AIC(\mathcal{M}) = n\{\log(2\pi) + \log(\text{SSE}(\mathcal{M})) - \log(n)\} + \frac{\text{SSE}(\mathcal{M})}{\sigma^2} + 2p(\mathcal{M}),$$

which is almost $C_p(\mathcal{M}) + K_n$.

The model with the lowest AIC is preferred.

e) Bayesian Information Criterion (BIC)

The BIC is formally defined as

$$BIC(\mathcal{M}) = -2\log L(\mathcal{M}) + p(\mathcal{M})log(n).$$

AIC assigns a less heavy penalty than BIC.