Chapter 2 Multiple Regression I  
(Part 1)

1 Regression several predictor variables

The response $Y$ depends on several predictor variables $X_1, ..., X_p$

\[ Y \]  \[ X_1, \ X_2, \ ..., \ X_p \]

Observations (or Design)

<table>
<thead>
<tr>
<th>obs.</th>
<th>$Y$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Y_1$</td>
<td>$X_{11}$</td>
<td>$X_{12}$</td>
<td>$X_{1p}$</td>
</tr>
<tr>
<td>2</td>
<td>$Y_2$</td>
<td>$X_{21}$</td>
<td>$X_{22}$</td>
<td>$X_{2p}$</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>$Y_n$</td>
<td>$X_{n1}$</td>
<td>$X_{n2}$</td>
<td>$X_{np}$</td>
</tr>
</tbody>
</table>

Thus, generally for individual $i$,

- the response is: $Y_i$
- the predictors variables are: $X_{i1}, X_{i2}, ..., X_{ip}$

2 Linear regression model with Two predictor variables

The linear regression model assumes that for any subject/individual with response $Y_i$ and predictor $X_{i1}, X_{i2}$ satisfies

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \]

where $E\varepsilon_i = 0$, or equivalently

\[ E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \]

Sometimes, it is also written as,

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \]
where $E\varepsilon = 0$. or equivalently

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

where $\beta_0, \beta_1, \beta_2$ are called regression coefficient

$\beta_0$ is called intercept

$\beta_1$ is called coefficient of $X_1$; $\beta_2$ is called coefficient of $X_2$

For example: (height in inch)

(Expected height of girl) = $-2.5 + 0.5($Farther’s height$) + 0.5($Mother’s height$)$

(Expected height of boy) = $2.5 + 0.5($Farther’s height$) + 0.5($Mother’s height$)$

**Meaning of the regression coefficients**

$\beta_1$ indicate the change in the mean response $EY$ per unit increase in $X_1$ when $X_2$ holds constant.

$\beta_2$ indicate the change in the mean response $EY$ per unit increase in $X_2$ when $X_1$ holds constant.

Note that $X_1$ and $X_2$ have some correlation, thus you need to know the difference in statistical and mathematical models [in mathematical model, $X_1$ and $X_2$ can be really free the change, but statistical model may not completely free]

### 3 Linear regression model with $p$ predictor variables

The linear regression model assumes that for any subject/individual with response $Y_i$ and predictor $X_{i1}, ..., X_{ip}$ satisfies

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \varepsilon_i$$

where $E\varepsilon_i = 0$, or equivalently

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$$

This means for each individual, the expected value of the response is a functional relationship with the independent variables. But the "real" value has random error $\varepsilon_i$ from the expected value.

Sometimes, the model is written as

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon$$
where $E\varepsilon = 0$, or equivalently

$$E(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

which is called a hyperplane, where $\beta_0, \beta_1, \ldots, \beta_p$ are called regression coefficients.

**Meaning of the regression coefficients**

$\beta_k$ indicate the change in the mean response $EY$ per unit increase in $X_k$ when the other predictors remain constant.

It is easy to see that we have studied the case $p = 1$, i.e. simple linear regression model.

We usually make the following assumptions

(L) Linearity (implied in the model)

(I) Independence of Error Terms, thus $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$, if $i \neq j$

(N) Normality of Error Terms: $\varepsilon \sim N(0, \sigma^2)$

(E) Equal/constant Error Variance: $\text{Var}\{\varepsilon_i\} = \sigma^2$

(F) Fixed design: $X_{i1}, ..., X_{ip}$ are known and nonrandom.

There are $p+1$ coefficients $\beta_0, ..., \beta_p$, one common variance $\sigma^2$, they are called parameters of the model.

**4 Some Examples**

Here we give some examples that are nonlinear, but can be transformed to linear regression models.

- Qualitative Predictor variables. It is understandable that the predictors must be quantitative. But we can also consider qualitative predictor, by denoting the predictor using dummy variables. For example $Y$ a person’s height, $X_1$ is his/her father’s height, $X_2$ his/her mother’s height, $S$ is the gender of the person. We can denote the gender by

  $$X_3 = \begin{cases} 
  1, & \text{if the person is male} \\
  0, & \text{if the person is male}
  \end{cases}$$

  Then our model is

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$
Polynomial regression models, for example

\[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i, \]
\[ Y_i = \beta_0 + \beta_1 X_i + \beta_3 X_i^2 + \ldots + \beta_k X_i^k + \varepsilon_i, \]
\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1}^2 + \beta_5 X_{i1} X_{i2} + \beta_6 X_{i3}^3 + \varepsilon_i, \]
\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1}^2 + \beta_5 X_{i1} X_{i2} + \beta_6 X_{i3}^4 + \varepsilon_i, \]

\(X_{i1}X_{i2}\) are usually called interaction of \(X_1\) and \(X_2\), how about \(X_{i2}X_{i3}\)?

- Transformed model (after variable transformation, the model become a linear regression model). Here are some examples

(a) For model \(Y_i = a_0 \exp(\beta_1 X_{i1} + \ldots + \beta_p X_{ip}) \xi_i\), let \(Z_i = \log(Y_i)\), \(\varepsilon_i = \log(\xi_i)\) and \(\beta_0 = \log(a_0)\). Taking logarithm, the model becomes

\[ Z_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i \]

(b) model \(Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1}^2 + \beta_5 X_{i1} X_{i2} + \beta_6 X_{i3}^3 + \varepsilon_i\), can be written as

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 Z_{i4} + \beta_5 Z_{i5} + \beta_6 Z_{i6} + \varepsilon_i, \]

where \(Z_{i4} = X_{i1}^2\), \(Z_{i5} = X_{i1} X_{i2}\) and \(Z_{i6} = X_{i3}^3\).

5 General linear regression model in matrix terms

Again, our general model can be written as

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, \ldots, n \]

or

\[ Y_1 = \beta_0 + \beta_1 X_{11} + \ldots + \beta_p X_{1p} + \varepsilon_1, \]
\[ Y_2 = \beta_0 + \beta_1 X_{21} + \ldots + \beta_p X_{2p} + \varepsilon_2, \]
\[ \vdots \]
\[ Y_n = \beta_0 + \beta_1 X_{n1} + \ldots + \beta_p X_{np} + \varepsilon_n \]

(with the 5 assumptions)
Let 
\[ X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}, \] called Design matrix.

\[ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \] called coefficient vector

\[ Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \] called response vector

\[ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \] called random error vector

It is easy to check
\[
\begin{bmatrix} 
1 & X_{11} & \cdots & X_{1p} \\
1 & X_{21} & \cdots & X_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
1 & X_{n1} & \cdots & X_{np} 
\end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_{11} + \cdots + \beta_p X_{1p} \\
\beta_0 + \beta_1 X_{21} + \cdots + \beta_p X_{2p} \\
\vdots \\
\beta_0 + \beta_1 X_{n1} + \cdots + \beta_p X_{np} \end{bmatrix}
\]

The regression model can be written as 
\[ \mathbf{Y} = \mathbf{X}\beta + \varepsilon \]

The (L-I-N-E) assumptions can be written as
\[ \mathbb{E}\{\varepsilon\} = 0, \quad \mathbb{V}\{\varepsilon\} = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I} \]
\[ \varepsilon \sim N(0, \sigma^2 \mathbf{I}) \]

6 Least squares estimation

- Minimize \[ Q(b_0, \ldots, b_p) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{i1} - \cdots - b_p X_{ip})^2 \]
- by calculus, we have the following \((p+1)\) Normal equations: (how?)
\[
\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{i1} - \cdots - b_p X_{ip}) = 0 \\
\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{i1} - \cdots - b_p X_{ip}) X_{i1} = 0 \\
\vdots \\
\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{i1} - \cdots - b_p X_{ip}) X_{ip} = 0
\]
• let \( b = (b_0, b_1, ..., b_p)' \). Then the Normal equations can be written as
\[ X'Xb = X'Y \]

• The solution, i.e. the estimator of the coefficient vector, is
\[ b = (X'X)^{-1}X'Y \]

• The estimated model is
\[ \hat{Y} = b_0 + b_1X_1 + ... + b_pX_p \]

• Fitted values
\[ \hat{Y}_i = b_0 + b_1X_{i1} + ... + b_pX_{ip}, \quad i = 1, ..., n \]

• (Fitted) residuals
\[ e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1X_{i1} + ... + b_pX_{ip}), \quad i = 1, ..., n \]

• Estimator of \( \sigma^2 \), denoted by \( \hat{\sigma}^2 \),
\[ MSE = \sum_{i=1}^{n} e_i^2 / \{n - (p + 1)\} \] called **Mean squared error**

why \((p+1)\)? (because there are \(p+1\) constraints, \(p+1\) is the number of (free) coefficients, or more exactly the number of Normal equations).

• **Dwaine Studios example** \( Y \)-sales, \( X_1 \)-number of persons aged 16 or less, \( X_2 \)-income. 21 observations

1. \[
Y = \begin{bmatrix} 174.4 \\ 164.4 \\ \vdots \\ 166.5 \end{bmatrix}; \quad X = \begin{bmatrix} 1 & 68.5 & 16.7 \\ 1 & 45.2 & 16.8 \\ \vdots & \vdots & \vdots \\ 1 & 52.3 & 16.0 \end{bmatrix}
\]

2. \[
X'X = \begin{bmatrix} 21.0 & 1,302.4 & 360.0 \\ 1,302.4 & 87,707.9 & 22,609.2 \\ 360.0 & 22,609.2 & 6,190.3 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 3,820 \\ 249,643 \\ 66,073 \end{bmatrix}
\]

3. \[
b = \begin{bmatrix} 21.0 & 1,302.4 & 360.0 \\ 1,302.4 & 87,707.9 & 22,609.2 \\ 360.0 & 22,609.2 & 6,190.3 \end{bmatrix}^{-1} \begin{bmatrix} 3,820 \\ 249,643 \\ 66,073 \end{bmatrix} = \begin{bmatrix} -68.85 \\ 1.45 \\ 9.37 \end{bmatrix}
\]
4. The estimated model is

\[ \hat{Y} = -68.85 + 1.45X_1 + 9.37X_2 \]

5. 

<table>
<thead>
<tr>
<th>obs.</th>
<th>(X_1)</th>
<th>(X_2)</th>
<th>(Y)</th>
<th>Fitted (\hat{Y}_i)</th>
<th>residuals (e_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68.5</td>
<td>16.7</td>
<td>174.4</td>
<td>187.184</td>
<td>-12.7841</td>
</tr>
<tr>
<td>2</td>
<td>45.2</td>
<td>16.8</td>
<td>164.4</td>
<td>154.229</td>
<td>10.1706</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>21</td>
<td>52.3</td>
<td>16.0</td>
<td>166.5</td>
<td>157.064</td>
<td>9.4356</td>
</tr>
</tbody>
</table>

6. 

\[ \hat{\sigma}^2 = MSE = \frac{\sum_{i=1}^{21} e_i^2}{n - p - 1} = \frac{2180.9274}{21 - 2 - 1} = 121.1626 \]

7. **Unbias of the estimators of coefficients**

The estimator of coefficient vector is unbiased, i.e.

\[ E(b) = \beta \]

and

\[ \text{Var}(b) = \sigma^2(X'X)^{-1} \]

In details

\[ E(b_k) = \beta_k \]

and

\[ \text{Var}(b_k) = \sigma^2c_{k+1,k+1}, \quad k = 0, 1, \ldots, p - 1 \]

where \(c_{kk}\) is the \((k, k)\)th entry in \((X'X)^{-1}\).

[Proof: Note that \(EY = X\beta\). Thus

\[ E\{b\} = E\{(X'X)^{-1}X'Y\} = (X'X)^{-1}X'E\{Y\} = \beta \]

and

\[ \text{Var}(b) = (X'X)^{-1}X'\text{Var}(Y)X(X'X)^{-1} = (X'X)^{-1}X'\sigma^2IX(X'X)^{-1} = \sigma^2(X'X)^{-1} \] ]
8 Fitted values and residuals in matrix form

- fitted value

\[
\hat{Y} = \begin{bmatrix}
\hat{Y}_1 \\
\hat{Y}_2 \\
\vdots \\
\hat{Y}_n \\
\end{bmatrix} = \begin{bmatrix}
b_0 + b_1 x_{11} + \ldots + b_p x_{1p} \\
b_0 + b_1 x_{21} + \ldots + b_p x_{2p} \\
\vdots \\
b_0 + b_1 x_{n1} + \ldots + b_p x_{np} \\
\end{bmatrix} = Xb
\]

- fitted residuals

\[
e = Y - \hat{Y} = (I - X(X'X)^{-1}X')Y
\]

Denote \( X(X'X)^{-1}X' \) by \( H \), we have \( \hat{Y} = HY, \ e = (I - H)Y \)

9 Variance-covariance matrix for residuals \( e \)

- \( \text{Var}\{e\} = \text{Var}\{(I - H)Y\} = (I - H)\text{Var}\{Y\}(I - H)' \)

- \( \text{Var}\{Y\} = \text{Var}\{ɛ\} = \sigma^2I \)

- \( (I - H)' = I' - H' = I - H \)

- \( HH = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = H \)

- \( (I - H)(I - H) = I - 2H + HH = I - H \)

- \( \text{Var}\{e\} = \sigma^2(I - H) \), which can be estimated by \( \hat{σ}^2(I - H) \), where

\[
\hat{σ}^2 = \text{MSE} = \frac{e'e}{n - p - 1} = \frac{Y'(I - H)Y}{(n - p - 1)}
\]

- \( E\hat{σ}^2 = E(MSE) = \sigma^2 \) [The proof can be ignored]
10 Variance-covariance matrix for $b$

Recall $b = (X'X)^{-1}X'Y$,

$$\text{Var}\{b\} = (X'X)^{-1}X'\text{Var}\{Y\}X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

where $\sigma^2$ can be estimated by $\hat{\sigma}^2 = \text{MSE}$. In other word, we estimate $\text{Var}\{b\}$ by $\hat{\sigma}^2(X'X)^{-1}$, denoted $s(b) = \hat{\sigma}^2(X'X)^{-1}$

For the above example,

$$\text{MSE} = \frac{\text{SSE}}{n - p - 1} = \frac{e'e}{21 - 2 - 1} = \frac{2,180.93}{18} = 121.16$$

$$s^2\{b\} = 121.16(X'X)^{-1} = \begin{bmatrix} 3,602.0 & 8.748 & -241.43 \\ 8.748 & 0.0448 & -0.679 \\ -241.43 & -0.679 & 16.514 \end{bmatrix}$$

11 The distribution of estimators

If $\mathcal{E} \sim N(0, \sigma^2 I)$ (i.e. $\varepsilon_i$ are IID $N(0, \sigma^2)$), then

- The estimated coefficients

$$b \sim N(\beta, \sigma^2(X'X)^{-1})$$

Denote the $(i, j)th$ entry of $(X'X)^{-1}$ by $c_{ij}$, then

$$b_k \sim N(\beta_k, \sigma^2 c_{k+1,k+1}), \quad k = 0, 1, ..., p - 1$$

(where $b = (b_0, b_1, ..., b_p)'$)

- Let $s(b_k) = \sqrt{\text{MSE} \ast c_{k+1,k+1}}$, called Standard Error (S.E.) for $b_k$ (which can be found in the output of R), then

$$\frac{b_k - \beta_k}{s(b_k)} \sim t(n - p - 1)$$

- $t$-value

$$t^* = \frac{b_k}{s(b_k)}$$
12 Confidence interval for $\beta_k$

with $1 - \alpha$ confidence, the Confidence interval for $\beta_k$ is

$$[b_k - s(b_k) * t(1 - \alpha/2, n - p - 1), \ b_k - s(b_k) * t(1 - \alpha/2, n - p - 1)]$$

For the Dwaine Studios example, the 95% Confidence interval for $\beta_2$ is

$$[9.3655 - 4.0640 \cdot 2.101, \ 9.3655 + 4.0640 \cdot 2.101] = [0.83, \ 17.90]$$

where quantile (critical value)

$$t(1 - \alpha/2, n - p - 1) = t(0.975, 21 - 3) = 2.101$$

is used

13 test for $\beta_k = 0$

Our hypothesis is

$$H_0 : \beta_k = 0, \ H_a : \beta_k \neq 0$$

under $H_0$,

$$t = \frac{b_k - \beta_k}{s(b_k)} = \frac{b_k}{s(b_k)} \sim t(n - p - 1)$$

For significant level $\alpha$, our criterion is

If the calculated $|t^*| > t(1 - \alpha/2, n - p - 1)$, reject $H_0$

If the calculated $|t^*| \leq t(1 - \alpha/2, n - p - 1)$, accept $H_0$

Similarly, we can do the test based on the p-value

If p-value < $\alpha$, reject $H_0$

If p-value $\geq \alpha$, accept $H_0$

For the Dwaine Studios example, test

$$H_0 : \beta_1 = 0, \ H_a : \beta_1 \neq 0$$

with significance level 5%, since

$$|t^*| = 6.868 > t(1 - \alpha/2, n - p - 1) = 2.101$$

we reject $H_0$. (in other words, $\beta_1$ is significantly different from 0.)

Similarly,

$H_0 : \beta_0 = 0$ can be accepted

$H_0 : \beta_2 = 0$ should be rejected
14 Prediction

For any new individual with \( X_{new} = (x_1, ..., x_p)^\top \), the predict mean response is

\[
\hat{Y}_{new} = \mathcal{X}'b
\]

where

\[
\mathcal{X} = (1, x_1, ..., x_p)^\prime
\]

We have

\[
E\hat{Y}_{new} = EY_{new}
\]

Note that if normal errors are assumed, i.e. \( \varepsilon_i \) are IID \( N(0, \sigma^2) \), then

\[
\hat{Y}_{new} \sim N(EY_{new}, \mathcal{X}'(X'X)^{-1}\mathcal{X}\sigma^2)
\]

Let

\[
s^2(\hat{Y}_{new}) = \mathcal{X}'(X'X)^{-1}\mathcal{X}\hat{\sigma}^2 = \mathcal{X}'(X'X)^{-1}\mathcal{X} \cdot MSE
\]

We have

\[
\frac{\hat{Y}_{new} - EY_{new}}{s(\hat{Y}_{new})} \sim t(n - p - 1)
\]

With confidence 100(1 - \( \alpha \))%, the C.I. for \( E(Y_{new}) \) is

\[
[\hat{Y}_{new} - s(\hat{Y}_{new}) \times t(1 - \alpha/2, n - p - 1), \hat{Y}_{new} + s(\hat{Y}_{new}) \times t(1 - \alpha/2, n - p - 1)]
\]

What about the prediction interval (P.I.) for the value \( Y_{new} \)? With confidence 100(1 - \( \alpha \))%, the P.I. for \( Y_{new} \) is

\[
[\hat{Y}_{new} - s(pred) \times t(1 - \alpha/2, n - p - 1), \hat{Y}_{new} + s(pred) \times t(1 - \alpha/2, n - p - 1)]
\]

where

\[
s^2(pred) = MSE + s^2(\hat{Y}_{new}) = MSE\{1 + \mathcal{X}'(X'X)^{-1}\mathcal{X}\}\}
\]

15 R code

- \texttt{regression=lm(y ~ x_1 + x_2 + ... + x_p)}
  \texttt{summary(regression)}

- \texttt{Xnew = data.frame(x1=c(...), x2=c(...), ..., xp=c(...))}
  \texttt{predict(regression, Xnew, interval = "confidence"/"prediction", level=0.95)}