Regression 6

Multiple Linear Regression I

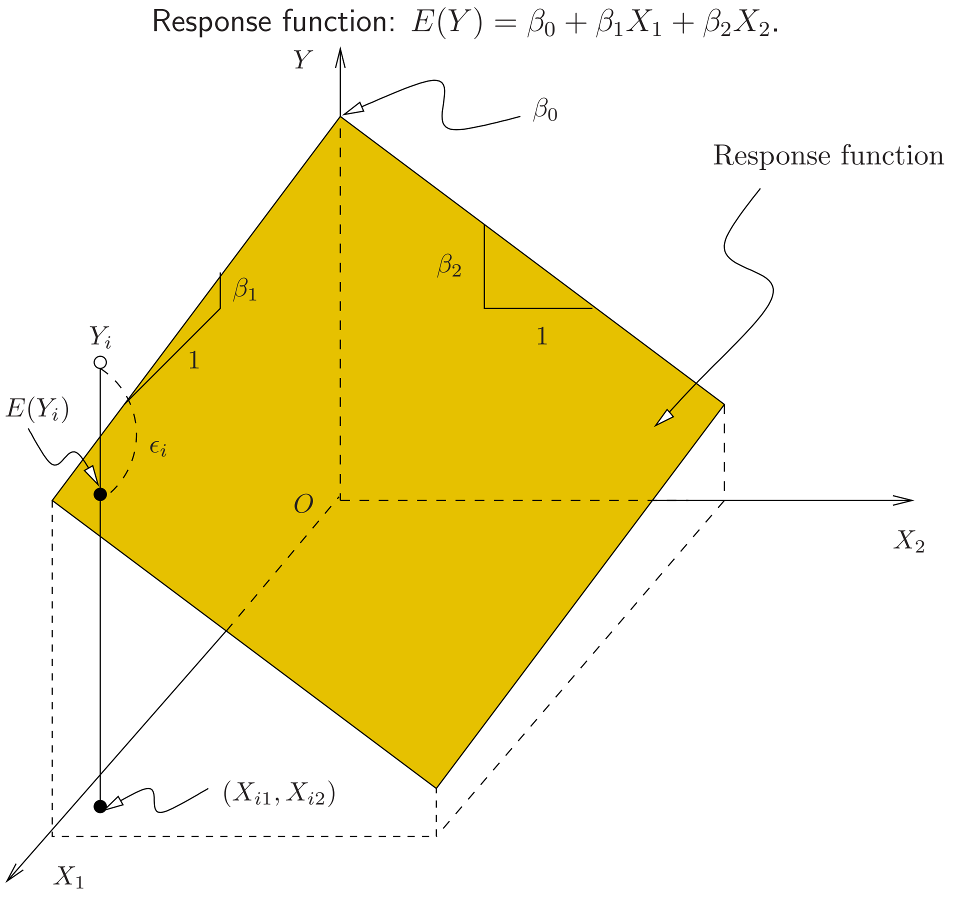
6.1 Introduction

In multiple linear regression, several predictors are used to model a single response variable.

When there $p-1$ predictors $(X_1, \ldots, X_{p-1})$, the linear regression model of a response Y_i is given by

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i \qquad (6.1) \\ &= \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \epsilon_i \end{aligned}$$

which is called a first-order model with $p-1$ predictors. We assume $E(\epsilon_i) = 0$ for $i = 1, \ldots, n$. A "first-order" model is linear in the predictors. Notice that we have p parameters $(\beta_0, \beta_1, \ldots, \beta_{p-1})$ with $p-1$ predictors.



- When $p = 2$, the equation (6.1) gives the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i.$$

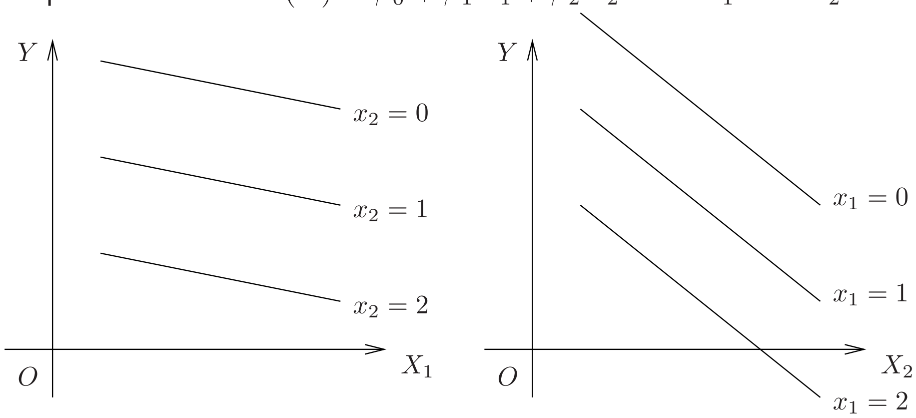
The regression function $E(Y) = \beta_0 + \beta_1 X_1$ is a line in the two-dimensional (X_1, Y) space.

- When $p = 3$, the equation (6.1) gives the two-predictor regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i.$$

The regression function $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ is a plane in the three-dimensional (X_1, X_2, Y) space.

Response function: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ with X_1 and X_2 fixed.



- When $p > 3$, the regression function

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}$$

is a *hyperplane* ($p-1$ dimensional plane) in the p-dimensional $(X_1, X_2, \ldots, X_{p-1}, Y)$ space.

6.2 General linear regression model

In general, the predictors $X_1, \ldots, X_{p-1}$ in a regression model do not need to represent different predictors. We define the general linear regression model, with normal error terms, simply in terms of the predictors $X_1, \ldots, X_{p-1}$.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \qquad i = 1, \ldots, n \ (n \geq p) \\ &= \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \epsilon_i \end{aligned}$$

where

1. $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are parameters

2. $X_{i1}, \ldots, X_{i,p-1}$ are known

3. ϵ_i are *iid* $N(0, \sigma^2)$, *i.e.*, $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

What is the difference between a linear model and non-linear model?

> *A linear model is defined as a model that is linear in the parameters, i.e., linear in the coefficients* $\beta_0, \beta_1, \ldots, \beta_{p-1}$.

This general linear model includes a variety of situations. We consider a few of them.

1. $p-1$ *predictor variables*

   The typical general linear regression model includes $p-1$ different *quantitative* predictor variables $X_1, \ldots, X_{p-1}$ with p parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$.

2. *Categorical predictor variables.*

   A very important application of regression analysis involves a list of predictors that includes *categorical variables* as well as usual traditional quantitative variables. The categorical variables are also called *indicator* variables or qualitative variables. For more details, see Chapter 11 of the textbook.

   For example, consider a regression analysis to predict the salary from gender (Z) and years employed (X). Let Z be defined as follows:

$$Z = \begin{cases} 1 & : \text{ if male} \\ 0 & : \text{ if female} \end{cases} .$$

   Thus, for the model

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \epsilon_i$$

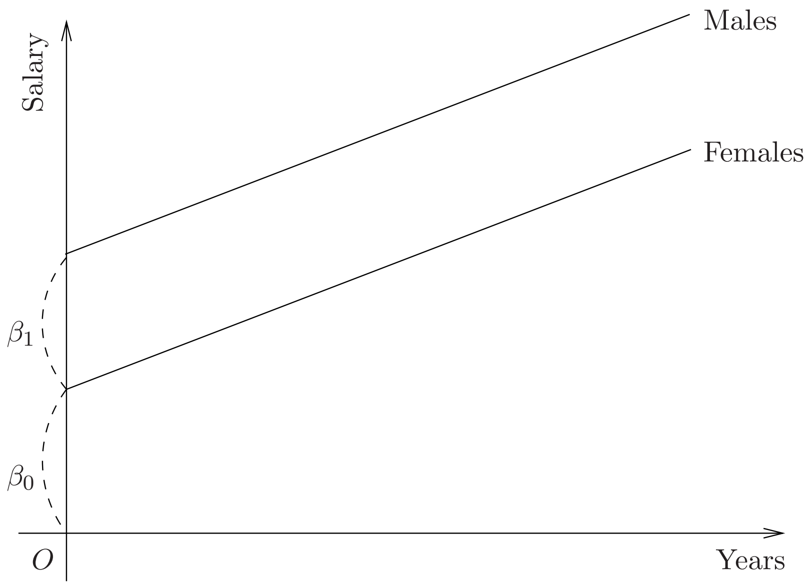
we have $\mathbf{X}$ matrix

$$\mathbf{X} = \left[\begin{array}{ccc} 1 & 0 & X_1 \\ 1 & 0 & X_2 \\ \vdots & \vdots & \vdots \\ \hline 1 & 1 & \vdots \\ \vdots & \vdots & \vdots \\ 1 & 1 & X_n \end{array}\right]$$

As a result, the model becomes:

$$Y_i = \begin{cases} (\beta_0 + \beta_1) + \beta_2 X_i + \epsilon_i & \text{male} \\ \beta_0 + \beta_2 X_i + \epsilon_i & \text{female} \end{cases}.$$

Categorical and continuous predictors



This categorical variable (gender) results in a mere *shift in intercept* induced by a constant different in response between the categories (male and female).

The role of the categorical variable, gender, can be determined by testing

$$H_0 : \beta_1 = 0 \ \text{ versus } \ H_1 : \beta_1 \neq 0.$$

If $\beta_1 \neq 0$, these two response functions represent parallel straight lines with different intercepts. The salary is a linear function of years employed (X) for both genders. The parameter β_1 indicates how much higher (or lower) the salary is than the one for males, for any given years employed.

Effect of years is same for both genders.

Effect of gender is same for all years.

$$\begin{aligned} \beta_0 &= \text{ intercept for females} \\ \beta_1 &= \text{ intercept for males } - \text{ intercept for females} \\ &= \text{ gender effect at any years} \\ \beta_2 &= \text{ slope for both} \end{aligned}$$

3. *Polynomial regression*

A polynomial regression model with one predictor variable

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \tag{6.2}$$

is also a special case of the general linear regression model despite the curvilinear nature of the response function (6.2). If we define

$$X_{i1} = X_i \ \text{ and } \ X_{i2} = X_i^2,$$

we can write (6.2) as follows:

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

which is in the form of the general linear regression model. Similarly models with higher-degree polynomial response functions are also particular cases of the general linear regression model.

4. *Transformed variables to Linearize.*

   Models with transformed variables are also special cases of the general linear regression model. For example the following model with a transformed Y:

$$\ln Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i.$$

   If we let $Y' = \ln Y_i$, we can write the above regression as follows

$$Y_i' = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i,$$

   which is in the form of general linear regression model (6.1).

   At this point we are curious about what kinds of examples fall into the category of non-linear models, *i.e.*, models not linear in the parameters. Given response Y and two regressors X_1 and X_2, the following represent two examples of non-linear models:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1^{\beta_3} + \beta_2 X_2^{\beta_4} + \epsilon \\ Y &= \frac{\beta_0}{1 + e^{-(\beta_1 X_1 + \beta_2 X_2)}} + \epsilon. \end{aligned}$$

5. *Interaction effects.*

   When the effects of the predictors on the response are *not additive*, the effect of one predictor variable depends on the levels of the other predictors. Note that a regression model with $p-1$ predictors contains *additive* effects if the response can be written in the form of:

$$Y = f_1(X_1) + f_2(X_2) + \cdots + f_{p-1}(X_{p-1}) + \epsilon, \tag{6.3}$$

where $f_1, f_2, \ldots, f_{p-1}$ can be any functions. For instance, the following response function with two predictors can be expressed in the form of:

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_1^2}_{f_1(X_1)} + \underbrace{\beta_3 X_2}_{f_2(X_2)} + \epsilon.$$

In contrast, the following regression function:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1 X_2 + \epsilon$$

cannot be expressed in the form of (6.3). Hence, this latter regression model is *not additive* (it contains an interaction effect).

A regression model to predict the salary from gender (Z) and years employed (X) with an added interaction term is

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \underbrace{\beta_3 X_i Z_i}_{\text{interaction}} + \epsilon_i.$$

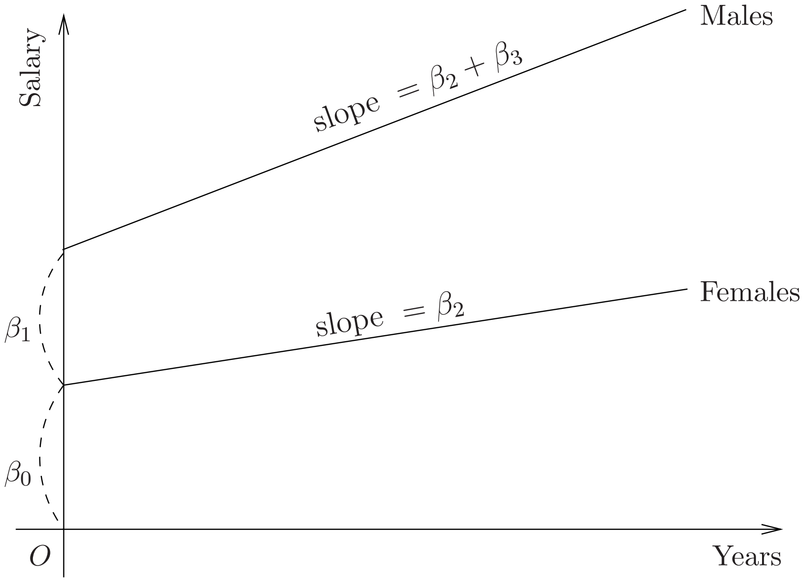
Let Z be defined as follows:

$$Z = \begin{cases} 1 & : \text{ if male} \\ 0 & : \text{ if female} \end{cases}.$$

As a result, the model becomes:

$$Y_i = \begin{cases} (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_i + \epsilon_i & \text{male} \\ \beta_0 + \beta_2 X_i + \epsilon_i & \text{female} \end{cases}.$$

Categorical and continuous predictors with interaction



$$\beta_0 = \text{ intercept for females}$$

$$\beta_1 = \text{ intercept for males } - \text{ intercept for females}$$

$$= \text{ gender effect "at years = 0"}$$

$$\beta_2 = \text{ slope for females}$$

$$= \text{ years effect for females}$$

$$\beta_3 = \text{ slope for males} - \text{ slope for females}$$

$$= \text{ years effect for females} - \text{ years effect for males}$$

$H_0 : \beta_3 = 0$ means lines are parallel

$H_0 : \beta_2 = 0$ means years has no effect for females

$H_0 : \beta_1 = 0$ means salaries for males and females are same at years $= 0$

6. *Combination of cases.*

   A regression model may combine several of the elements we have just noted and still be treated as a general linear regression model.

6.3 General linear regression model in matrix notation

The general linear regression model defined in (6.1) can be expressed in matrix notation. We need to define the following matrices:

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \underset{n\times p}{\mathbf{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \quad \underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \quad \underset{n\times 1}{\boldsymbol{\epsilon}} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

In matrix notation, the general regression model (6.1) becomes:

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}} \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}.$$

1. $\mathbf{Y}$ is a random vector which concerns the response variables.

2. $\mathbf{X}$ is a matrix of data which concerns the predictor variables.

   $\mathbf{X}$ is assumed known (fixed).

3. $\boldsymbol{\beta}$ is the parameter vector.

4. $\boldsymbol{\epsilon}$ is a random vector such that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.

Consequently, the random vector $\mathbf{Y}$ has the expectation and covariance matrix:

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad \mathrm{Cov}(\mathbf{Y}) = \mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I},$$

and it has a normal distribution

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}).$$

6.4 Estimation of regression parameters

1. *Least-squares method*

   The least squares criterion function Q_2 is generalized for general linear regression model:

   $$Q_2(\beta_0, \ldots, \beta_{p-1}) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1}X_{i,p-1})^2.$$

   The values of $\beta_0, \beta_1, \ldots, \beta_{p-1}$ that minimize Q_2 can be derived by differentiating the above Q_2 function with respect to $\beta_0, \beta_1, \ldots, \beta_{p-1}$. It gives the following normal equations:

   $$\begin{aligned}
   \frac{\partial Q_2}{\partial \beta_0} &= -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1}X_{i,p-1}) = 0 \\
   \frac{\partial Q_2}{\partial \beta_1} &= -2\sum_{i=1}^{n}X_{i1}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1}X_{i,p-1}) = 0 \\
   &\vdots \\
   \frac{\partial Q_2}{\partial \beta_{p-1}} &= -2\sum_{i=1}^{n}X_{i,p-1}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1}X_{i,p-1}) = 0.
   \end{aligned}$$

   Let $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1}]'$ be the solution for the p simultaneous equations. Then this solution is the least-squares estimates of the parameters $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_{p-1}]'$.

2. *Using the projection*

   It is very difficult to solve the p simultaneous normal equations above for $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$. It is easier to use the geometry of regression of $\mathbf{Y}$ onto the space $\mathcal{R}(\mathbf{1}, \mathbf{X}_1, \ldots, \mathbf{X}_{p-1})$, where $\mathbf{1} = [1, 1, \ldots, 1]'$ and $\mathbf{X}_k = [X_{1k}, X_{2k}, \ldots, X_{nk}]'$ for $k = 1, \ldots, p-1$. Using the projection idea, we have the least squares estimators for $\boldsymbol{\beta}$, which are

$$\underset{p\times 1}{\hat{\boldsymbol{\beta}}} = \underbrace{(\mathbf{X}'\mathbf{X})}_{p\times p}{}^{-1}\underbrace{(\mathbf{X}'\mathbf{Y})}_{p\times 1}.$$

3. *MLE*

   The likelihood function for the general linear regression with normal error $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is

$$\begin{aligned}
L(\boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{n} f(\epsilon_i) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big[-\frac{1}{2\sigma^2}\epsilon_i^2\Big] \\
&= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\Big[-\frac{1}{2\sigma^2}\epsilon_1^2 - \cdots - \frac{1}{2\sigma^2}\epsilon_n^2\Big] \\
&= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\Big[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2\Big] \\
&= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\Big[-\frac{1}{2\sigma^2}Q_2\Big].
\end{aligned}$$

   Maximizing this likelihood function with respect to $\beta_0, \beta_1, \ldots, \beta_{p-1}$ leads to

$$\underset{p\times 1}{\hat{\boldsymbol{\beta}}} = \underbrace{(\mathbf{X}'\mathbf{X})}_{p\times p}{}^{-1}\underbrace{(\mathbf{X}'\mathbf{Y})}_{p\times 1}.$$

   These estimators are least-squares and MLE.

6.5 Fitted values and residuals

The fitted values $\hat{Y}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_{p-1} X_{i,p-1}, \qquad i = 1, \ldots, n.$$

In matrix notation, we have

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

where $\hat{\mathbf{Y}} = [\hat{Y}_1, \ldots, \hat{Y}_n]'$ and $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1}]'$. Since $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, we have

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

$$\boxed{\hat{\mathbf{Y}} = \overbrace{\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}^{\mathbf{H}}\,\mathbf{Y}, \quad \mathbf{X}\,\underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbf{Y}}_{\hat{\boldsymbol{\beta}}}}$$

Let the vector of the residuals $\hat{\epsilon}_i = Y_i - \hat{Y}_i$ be denoted by

$$\underset{n\times 1}{\hat{\boldsymbol{\epsilon}}} = \begin{bmatrix} \hat{\epsilon}_1 \\ \vdots \\ \hat{\epsilon}_n \end{bmatrix}.$$

Then we have

$$\underset{n\times 1}{\hat{\boldsymbol{\epsilon}}} = \underset{n\times 1}{\mathbf{Y}} - \underset{n\times 1}{\hat{\mathbf{Y}}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}.$$

The expectation of the vector of residuals $\hat{\boldsymbol{\epsilon}}$ is

$$E\big[\hat{\boldsymbol{\epsilon}}\big] = E\big[(\mathbf{I} - \mathbf{H})\mathbf{Y}\big] = (\mathbf{I} - \mathbf{H})E[\mathbf{Y}] = (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = (\mathbf{X} - \mathbf{H}\mathbf{X})\boldsymbol{\beta} = (\mathbf{X} - \mathbf{X})\boldsymbol{\beta} = \mathbf{0}.$$

The covariance matrix of the vector of residuals $\hat{\boldsymbol{\epsilon}}$ is

$$\mathrm{Cov}(\hat{\boldsymbol{\epsilon}}) = \mathrm{Cov}((\mathbf{I}-\mathbf{H})\mathbf{Y}) = (\mathbf{I}-\mathbf{H})\,\mathrm{Cov}(\mathbf{Y})(\mathbf{I}-\mathbf{H})' = \sigma^2(\mathbf{I}-\mathbf{H}),$$

and is estimated by

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\epsilon}}) = \mathrm{MSE}\,(\mathbf{I}-\mathbf{H}),$$

where

$$\mathrm{MSE} = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2.$$

6.6 ANOVA results

6.6.1 ANOVA table

The sums of squares can be expressed in matrix notation.

$$\begin{aligned}
\mathrm{SSTO} &= \sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \|\mathbf{Y}-\bar{Y}\mathbf{1}\|^2 \\
&= \sum_{i=1}^{n}Y_i^2 - \frac{1}{n}(\sum_{i=1}^{n}Y_i)^2 = \mathbf{Y}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y} \\
\mathrm{SSE} &= \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 = \|\mathbf{Y}-\hat{\mathbf{Y}}\|^2 \\
&= (\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b}) = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} \\
\mathrm{SSR} &= \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 = \|\hat{\mathbf{Y}}-\bar{Y}\mathbf{1}\|^2 \\
&= \mathrm{SSTO} - \mathrm{SSE} \\
&= \mathbf{b}'\mathbf{X}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y}
\end{aligned}$$

Projecting the vector $\mathbf{Y}$ onto $\mathbf{1}$, we have the same result:

$$\|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 + \|\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}\|^2$$

$$\text{SSTO} \quad = \quad \text{SSE} \; + \; \text{SSR}$$

ANOVA table

Source	SS	df	MS	F
Regression	SSR	$p-1$	MSR = SSR/$(p-1)$	$F = \frac{\text{MSR}}{\text{MSE}}$
Error	SSE	$n-p$	MSE = SSE/$(n-p)$	
Total	SSTO	$n-1$		

6.6.2 Overall F-test

The F statistic in the ANOVA table is used for the overall (or omnibus) F-test which tests the significance of all predictors at once:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

$$H_1 : \text{not all } \beta_j = 0, (j = 1, \ldots, p-1).$$

The above test is equivalent to testing

$$H_0 : \boldsymbol{\beta^*} = \mathbf{0} \;\; \text{versus} \;\; H_1 : \boldsymbol{\beta^*} \neq \mathbf{0},$$

where $\boldsymbol{\beta^*} = [\beta_1, \beta_2, \cdots, \beta_{p-1}]'$.

The decision rule at the significance level α is :

$$\text{If } F \leq F(1-\alpha; p-1, n-p), \quad \text{conclude } H_0$$

$$\text{If } F > F(1-\alpha; p-1, n-p), \quad \text{conclude } H_1.$$

Notes about overall F-test:

1. If H_0 is rejected, we know that at least one of the coefficients $\beta_1, \ldots, \beta_{p-1}$ is non-zero. But we don't know which one(s) is/are non-zero.

2. Suppose that only a few predictors are important but the rest are not, *i.e.*, that only a few of the coefficients $\beta_1, \ldots, \beta_{p-1}$ are non-zero. Then we might fail to reject to H_0 because the significance of the important predictors is watered down by the unimportant ones. Note that we have

$$F = \left(\frac{n-p}{p-1}\right)\frac{\mathrm{SSR}}{\mathrm{SSE}}.$$

As p gets larger, SSR tends to increase but $\left(\frac{n-p}{p-1}\right)$ tends to decrease. Thus, when p is large, a watered-down effect can be made.

6.6.3 Coefficient of multiple determination

The *coefficient of multiple determination* R^2 is defined as

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SSTO}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SSTO}}$$

It measures the proportion of variance of Y explained by $X_1, \ldots, X_{p-1}$.

Notes about R^2:

1. The *coefficient of multiple correlation* R is the positive square root of R^2:

$$R = \sqrt{R^2}.$$

2. When $p = 2$, the coefficient of multiple correlation R is equal to the absolute value of the sample correlation coefficient r. (*i.e.*, $R = |r|$ when $p = 2$).

3. It can be shown that the coefficient of multiple determination R^2 can be viewed as the a coefficient of simple determination between the responses Y_i and the fitted values $\hat{Y}_i$ (or, a squared sample correlation between Y_i and $\hat{Y}_i$).

4. It can be shown that
$$R^2 = \frac{F}{F + (n-p)/(p-1)},$$
where $F = \text{MSR}/\text{MSE} = [\text{SSR}/(p-1)]/[\text{SSE}/(n-p)]$.

5. A large value of R^2 does not necessarily imply that the fitted model is a useful one.

6. Adding more predictors (X) to the regression model always increases R^2 (equivalently, this decreases SSE). Since R^2 usually can be made larger by including a larger number of predictors, it is sometimes suggested that a modified measure be used that adjusts for the number of X variables in the model. The *adjusted coefficient of multiple determination*, denoted by R^2_{adj}, adjust R^2 by dividing each sum of squares by its associated degrees of freedom:
$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-p)}{\text{SSTO}/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)\frac{\text{SSE}}{\text{SSTO}} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2).$$

6.7 Inferences in regression analysis

1. *Regression coefficient.*

   The least squares estimators and MLE in $\hat{\boldsymbol{\beta}}$ are unbiased, *i.e.*, $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

$$\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \mathrm{Cov}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{Cov}(\mathbf{Y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\
\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) &= \mathrm{MSE}\,(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

   The estimators in $\hat{\boldsymbol{\beta}}$ are distributed as

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

2. *Interval estimation of β_j.*

   For the normal error regression model, we have

$$\frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}(\hat{\beta}_j)} \sim t(\mathrm{df} = n - p), \qquad j = 0, 1, 2, \ldots, p-1.$$

   The $\left[\mathrm{SE}(\hat{\beta}_j)\right]^2$ can be obtained from the corresponding diagonal element of $\mathrm{MSE} \cdot (\mathbf{X}'\mathbf{X})^{-1}$. Hence, the confidence limits for β_j with $1-\alpha$ confidence coefficient are:

$$\hat{\beta}_j \pm t(1 - \frac{\alpha}{2}; n-p) \cdot \mathrm{SE}(\hat{\beta}_j).$$

3. *Test for β_j.*

   To test

$$H_0 : \beta_j = 0 \ \text{ versus } \ H_1 : \beta_j \neq 0,$$

we can use the test statistic:

$$T = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)}$$

and the decision rule:

$$\text{if } |T| \le t(1-\frac{\alpha}{2}; n-p), \quad \text{conclude } H_0$$

$$\text{Otherwise conclude } H_1.$$

4. *Bonferroni joint confidence intervals.*

   If g parameters are to be estimated jointly (where $g \le p$), the confidence limits with family confidence coefficient $1-\alpha$ are:

$$\hat{\beta}_j \pm t\big(1-\frac{\alpha}{2g}; n-p\big) \cdot \mathrm{SE}(\hat{\beta}_j).$$

5. *Mean response at X_h.*

   Define $\mathbf{x}_h' = \begin{bmatrix}1 & X_{h1} & X_{h2} & \ldots & X_{h,p-1}\end{bmatrix}$. Then the fitted value at $X_{h1}, X_{h2}, \ldots, X_{h,p-1}$ in matrix notation is

$$\hat{Y}_h = \mathbf{x}_h'\hat{\boldsymbol{\beta}}.$$

   Hence we have

$$\begin{aligned}
\mathrm{Var}(\hat{Y}_h) &= \mathrm{Cov}(\hat{Y}_h) = \mathrm{Cov}\left(\mathbf{x}_h'\hat{\boldsymbol{\beta}}\right) \\
&= \mathbf{x}_h' \,\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}\right)\mathbf{x}_h \\
&= \mathbf{x}_h'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h \\
&= \sigma^2 \cdot \mathbf{x}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h \\
\widehat{\mathrm{Var}}(\hat{Y}_h) &= \mathrm{MSE} \cdot \mathbf{x}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h.
\end{aligned}$$

6. *Prediction of new observation,* $Y_{h(\text{new})}$

   Recall that we have studied the following in §2.4 and §5.10:

$$\mathrm{Var}(Y_{h(\text{new})} - \hat{Y}_h) = \mathrm{Var}(Y_{h(\text{new})}) + \mathrm{Var}(\hat{Y}_h)$$

   Thus, the above variance in matrix notation becomes

$$\mathrm{Var}(Y_{h(\text{new})} - \hat{Y}_h) = \sigma^2\big\{1 + \mathbf{x}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h\big\}.$$

   Hence, we have

$$\widehat{\mathrm{Var}}(Y_{h(\text{new})} - \hat{Y}_h) = \mathrm{MSE}\big\{1 + \mathbf{x}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h\big\}.$$

6.8 Box-Cox transformations for multiple regression models

The updated Box-Cox programs (`BoxCox.MAC` and `BoxCox.R`) can be downloaded at:

`https://github.com/AppliedStat/LM`

These updated versions can support any linear regression models, while the older version can handle only a simple regression.

`Minitab`

For example, suppose that `c1`, `c2` and `c3` are predictors, `c4` is a response variable, and `c5` is a sequence of λ for the Box-Cox transformation. Then `BoxCox.MAC` `Minitab` macro function at `https://github.com/AppliedStat/LM-mtb` calculates `MSE` values (default option) and saves them onto `c6`. If `SSE` is preferred, then use the `SSE` subcommand as below.

```
MTB> set c5
MTB>   -10:10
MTB> end
MTB> let  c5 = c5/10

MTB> %U:\math8050\minitab\BOXCOX c4 c1-c3 c5 c6;
SUBC>  SSE.
```

R

Suppose that `x1`, `x2` and `x3` are predictors, `y` is a response variable, and `lam` is a sequence of λ for the Box-Cox transformation. Then `BoxCox` R function calculates `MSE` values (default option). If `SSE` is preferred, then use the `SSE=TRUE` option as below.

```
> source("https://raw.githubusercontent.com/AppliedStat/LM/master/BoxCox.R")
> lam = seq(-1, 1, 0.1)
> SSE = BoxCox ( y ~ x1 + x2 + x3, lambda=lam, SSE=TRUE )
> plot(lam, SSE)
```

6.9 Example: Patient-satisfaction data

We shall develop a multiple regression application with three predictors. We will analyze the Problems 6.15 $\sim$ 6.17 on Page 251 of the textbook. Section 6.9 of the textbook also provides a very good example. A hospital administrator wished to study the relation between

Y : Patient satisfaction,

X_1: Patient's age in years,

X_2: Severity of illness (an index),

X_3: Anxiety level (an index).

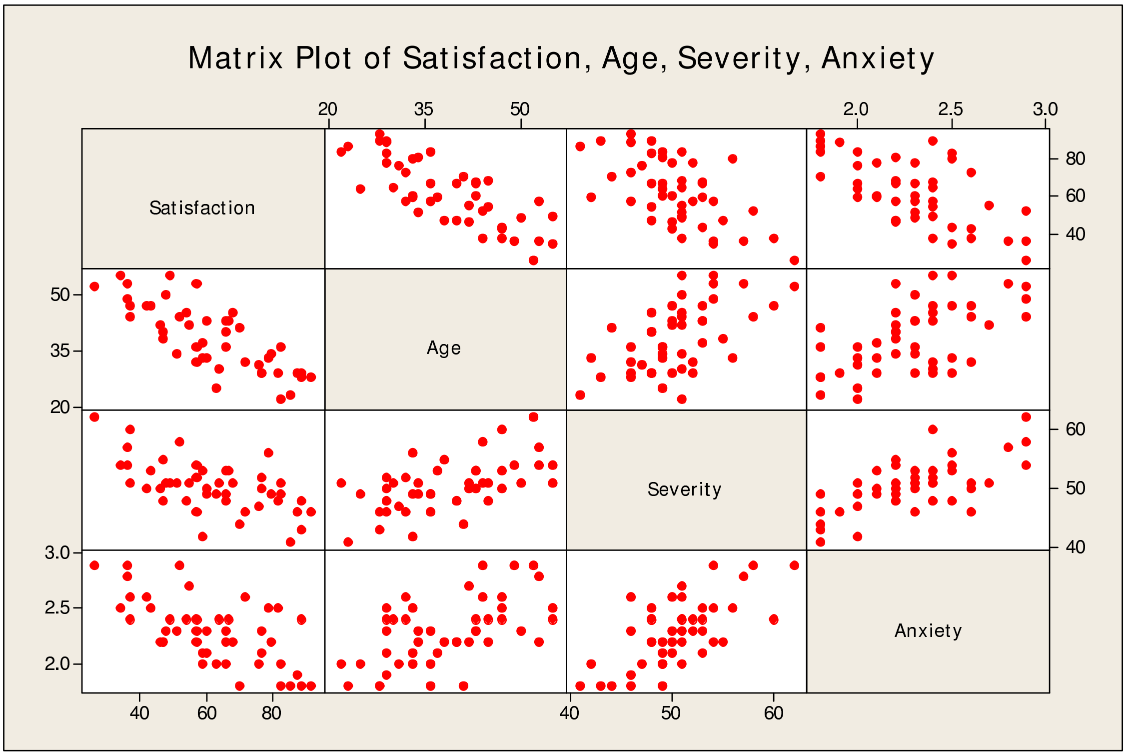
Minitab

1. *Read the data.*

```
MTB >  read c4 c1-c3;
SUBC>       file "U:\math8050\data\CH06PR15.TXT" .
Entering data from file: U:\MATH8050\DATA\CH06PR15.TXT
46 rows read.
MTB >  NAME c1  "Age"
MTB >  NAME c2  "Severity"
MTB >  NAME c3  "Anxiety"
MTB >  NAME c4  "Satisfaction"
```

2. *Scatter plot matrix.*

```
MTB >  matrixplot c4 c1-c3
Matrix Plot of Satisfaction, Age, Severity, Anxiety
```



3. *Correlation matrix.*

```
MTB >  correlation c4 c1-c3
Correlations: Satisfaction, Age, Severity, Anxiety

              Satisfaction          Age      Severity
Age                 -0.787
                     0.000

Severity            -0.603        0.568
                     0.000        0.000

Anxiety             -0.645        0.570         0.671
                     0.000        0.000         0.000

Cell Contents: Pearson correlation
               P-Value
```

4. *Regression of Y on X_1, X_2, and X_3.*

```
MTB >  regr c4 3 c1 c2 c3;
SUBC>  fits  c5;
SUBC>  resid c6.

Regression Analysis: Satisfaction versus Age, Severity, Anxiety

The regression equation is
Satisfaction = 158 - 1.14 Age - 0.442 Severity - 13.5 Anxiety

Predictor     Coef  SE Coef      T      P
Constant    158.49    18.13   8.74  0.000
Age        -1.1416   0.2148  -5.31  0.000
Severity   -0.4420   0.4920  -0.90  0.374
Anxiety    -13.470    7.100  -1.90  0.065

S = 10.0580   R-Sq = 68.2%   R-Sq(adj) = 65.9%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       3   9120.5  3040.2  30.05  0.000
Residual Error  42   4248.8   101.2
Total           45  13369.3

Source    DF  Seq SS
Age        1  8275.4
Severity   1   480.9
Anxiety    1   364.2
```

5. *90% Confidence interval of the mean response and prediction interval when $X_1 = 35$, $X_2 = 45$, $X_3 = 2.2$.*

```
MTB >  regress c4 3 c1-c3;
SUBC>    predict 35 45 2.2;
SUBC>    confidence 90.

Predicted Values for New Observations

New
Obs    Fit  SE Fit      90% CI          90% PI
  1  69.01    2.66  (64.53, 73.49)  (51.51, 86.51)
```

6. *Tests for constancy of error variance.*

   Non-constancy of variance of error can be detected by $\hat{\epsilon}_i^2$ versus $\hat{Y}_i$ plot. The modified Levene test and the *Breusch-Pagan test* are two typical tests for constancy of error variance. Here we present the Breusch-Pagan test. This test assumes that the error terms are independent and normally distributed and the variance of the error term ϵ_i, denoted by σ_i^2 is related to the levels of $X_1, \ldots, X_{p-1}$ in the following way:

$$\ln \sigma_i^2 = \gamma_0 + \gamma_1 X_{i1} + \cdots + \gamma_{p-1} X_{i,p-1}.$$

   The test of $H_0 : \gamma_1 = \cdots = \gamma_{p-1} = 0$ is carried out by means of regressing the squared residuals $\hat{\epsilon}_i^2$ on $X_1, \ldots, X_{p-1}$ in the usual manner and obtaining the regression sum of squares SSR*. The test statistic X^2_{BP} is as follows:

$$X^2_{\text{BP}} = \frac{\text{SSR}^*}{2} \div \left(\frac{\text{SSE}}{n}\right)^2 \mathrel{\dot\sim} \chi^2(df = p-1),$$

   where SSR* is the regression sum of squares when regressing $\hat{\epsilon}_i^2$ on $X_1, \ldots, X_{p-1}$ and SSE is the error sum of squares when regressing Y on $X_1, \ldots, X_{p-1}$. Large values of X^2_{BP} lead to H_1: non-constancy of error variance.

```
MTB >  let c22=c6*c6
MTB >  regr c22 3 c1-c3

Regression Analysis: C22 versus Age, Severity, Anxiety

The regression equation is
C22 = 49 - 2.73 Age + 3.81 Severity - 19.4 Anxiety

Predictor    Coef  SE Coef      T      P
Constant     49.4    167.5   0.30  0.769
Age        -2.728    1.984  -1.37  0.176
Severity    3.807    4.545   0.84  0.407
Anxiety    -19.38    65.59  -0.30  0.769

S = 92.9193   R-Sq = 5.6%   R-Sq(adj) = 0.0%

Analysis of Variance
```

```
Source          DF      SS    MS     F      P
Regression       3   21356  7119  0.82  0.488
Residual Error  42  362628  8634
Total           45  383983

Source    DF  Seq SS
Age        1   15085
Severity   1    5517
Anxiety    1     754
```

From the `Minitab` results, we have the test statistic

$$X^2_{\text{BP}} = \frac{21356}{2} \div \left(\frac{4248.8}{46}\right)^2 = 1.25.$$

If we use the significance level $\alpha = 0.01$, we have the critical value $\chi^2(0.99; 3) = 11.34$.

Comparing $X^2_{\text{BP}} = 1.25$ with $\chi^2(0.99; 3) = 11.34$, we conclude that error variance is constant.

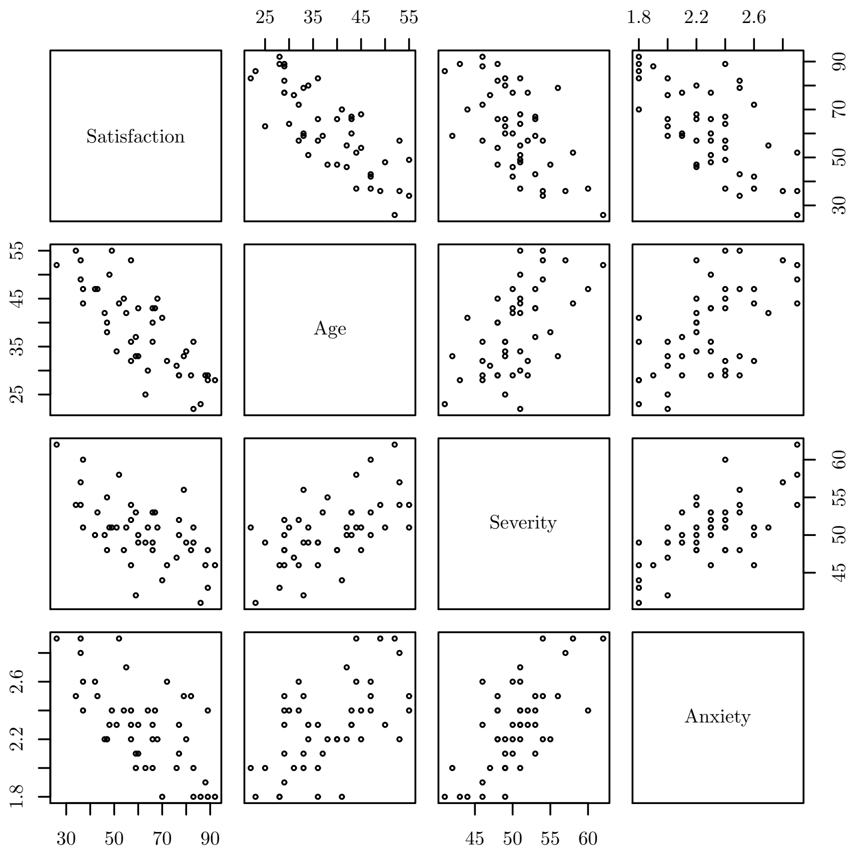
The `Minitab` macro for the Breusch-Pagan test (file: `BPtest.MAC`) is also available at `https://github.com/AppliedStat/LM-mtb`

```
MTB >  read c4 c1-c3;
SUBC>   file "U:\math8050\data\CH06PR15.TXT" .
Entering data from file: U:\MATH8050\DATA\CH06PR15.TXT
46 rows read.

MTB >  %U:\math8050\minitab\BPtest c4 c1-c3 .
Executing from file: U:\math8050\minitab\BPtest.MAC

Data Display

Breusch-Pagan Test Statistic:      1.25157
Degrees of Freedom:                3
p-value:                           0.74066
```

R

1. *Read the data.*

```
> mydata=read.table("https://raw.githubusercontent.com/AppliedStat/LM/master/CH06PR15.txt")
> y  = mydata[,1]
> x1 = mydata[,2]
> x2 = mydata[,3]
> x3 = mydata[,4]
```



2. *Scatter plot matrix.*

```
> colnames(mydata) = c("Satisfaction", "Age", "Severity", "Anxiety")
> pairs (mydata, cex=0.5, pch=1)
```

3. *Correlation matrix.*

```
>  cor( mydata )
             Satisfaction        Age   Severity    Anxiety
Satisfaction    1.0000000 -0.7867555 -0.6029417 -0.6445910
Age            -0.7867555  1.0000000  0.5679505  0.5696775
Severity       -0.6029417  0.5679505  1.0000000  0.6705287
Anxiety        -0.6445910  0.5696775  0.6705287  1.0000000
```

4. *Regression of Y on X_1, X_2, and X_3.*

```
>  LM = lm ( y ~ x1 + x2 + x3 )
>
>  summary(LM)
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-18.3524  -6.4230   0.5196   8.3715  17.1601

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
x1           -1.1416     0.2148  -5.315 3.81e-06 ***
x2           -0.4420     0.4920  -0.898   0.3741
x3          -13.4702     7.0997  -1.897   0.0647 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.06 on 42 degrees of freedom
Multiple R-Squared: 0.6822,     Adjusted R-squared: 0.6595
F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10

>  anova (LM)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 8275.4  8275.4 81.8026 2.059e-11 ***
x2         1  480.9   480.9  4.7539   0.03489 *
x3         1  364.2   364.2  3.5997   0.06468 .
Residuals 42 4248.8   101.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. *90% Confidence interval of the mean response and prediction interval when $X_1 = 35$, $X_2 = 45$, $X_3 = 2.2$.*

```
>  new = data.frame( x1=35, x2=45, x3=2.2 )
>  predict ( LM, newdata=new, interval="confidence", level=0.90)
          fit      lwr      upr
[1,] 69.01029 64.52854 73.49204
>
>  new = data.frame( x1=35, x2=45, x3=2.2 )
>  predict ( LM, newdata=new, interval="prediction", level=0.90)
          fit      lwr      upr
[1,] 69.01029 51.50965 86.51092
```

6. *Tests for constancy of error variance.*
   We present the Breusch-Pagan test using R.

```
>  e = resid(LM)
>  SSE = sum( e^2 )
>  sigma2 = e^2
>
>  LM2 = lm ( sigma2 ~ x1 + x2 + x3 )
>  SSR.star = sum( (fitted(LM2)-mean(sigma2))^2 )
>
```

```
>  n = length(y)
>
>  cbind(SSR.star, SSE, n)
     SSR.star      SSE  n
[1,] 21355.53 4248.841 46
>
>  X.BP = SSR.star/2 / ( (SSE/n)^2 )
>
>  X.BP
[1] 1.251570
>
>  qchisq(0.99, df = 3) ## chi-square critical value
[1] 11.34487
```

The R function for the Breusch-Pagan test (file: `Breusch-Pagan.R`) is also available at

https://github.com/AppliedStat/LM

```
> # 1. Read the data
>
> mydata=read.table("https://raw.githubusercontent.com/AppliedStat/LM/master/CH06PR15.txt")
>
> y  = mydata[,1]
> x1 = mydata[,2]
> x2 = mydata[,3]
> x3 = mydata[,4]
>
>
> # 2. Read Breusch-Pagan test R program
>
> source("https://raw.githubusercontent.com/AppliedStat/LM/master/Breusch-Pagan.R")
>
> BP.test ( y ~ x1 + x2 + x3)
$test.stat
[1] 1.251570

$df
[1] 3

$p.value
[1] 0.7406642
```