

Multiple Linear Regression I

1 Introduction

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

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

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

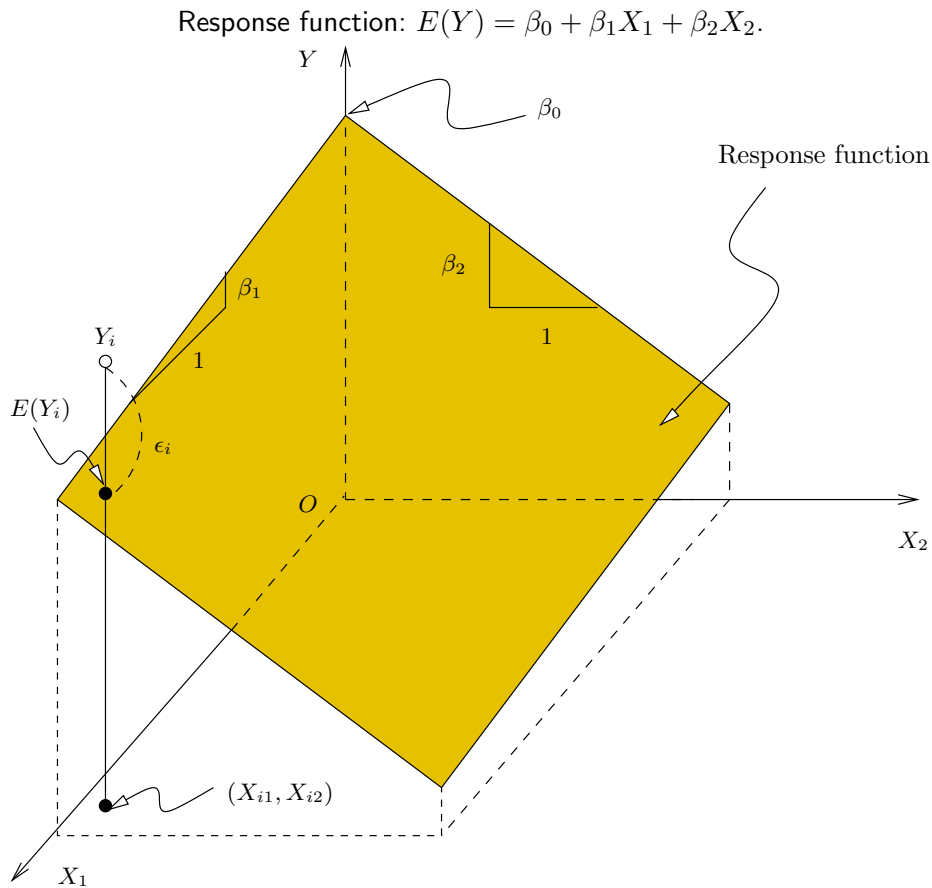
which is called a first-order model with $p - 1$ predictors. We assume $E(\epsilon_i) = 0$ for $i = 1, \dots, n$. A “first-order” model is linear in the predictors. Notice that we have p parameters $(\beta_0, \beta_1, \dots, \beta_{p-1})$ with $p - 1$ predictors.

Remark 6.1.

- 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.

- When $p > 3$, the regression function

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

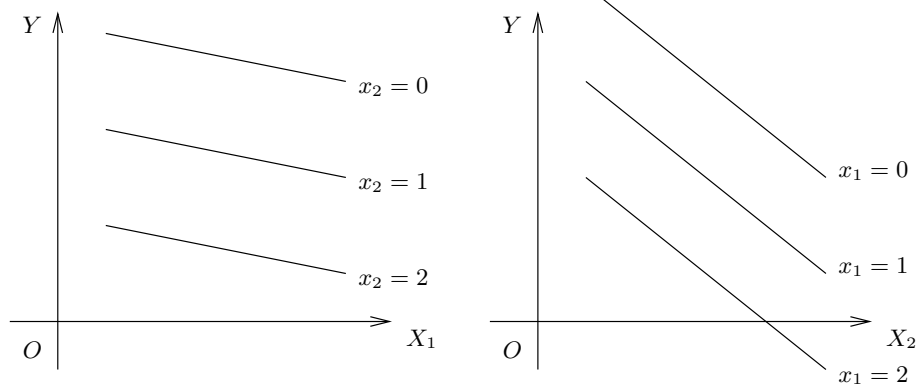
is a *hyperplane* ($p - 1$ dimensional plane) in the p -dimensional $(X_1, X_2, \dots, X_{p-1}, Y)$ space.

△

2 General linear regression model

In general, the predictors X_1, \dots, 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,

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



simply in terms of the predictors X_1, \dots, X_{p-1} .

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \quad i = 1, \dots, n \quad (n \geq p)$$

$$= \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \epsilon_i$$

where

1. $\beta_0, \beta_1, \dots, \beta_{p-1}$ are parameters
2. $X_{i1}, \dots, X_{i,p-1}$ are known
3. ϵ_i are iid $N(0, \sigma^2)$, i.e., $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$.

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 $\beta_0, \beta_1, \dots, \beta_{p-1}$.

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, \dots, X_{p-1} with p parameters $\beta_0, \beta_1, \dots, \beta_{p-1}$.

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

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & X_1 \\ 1 & 0 & X_2 \\ \vdots & \vdots & \vdots \\ 1 & 1 & \vdots \\ \vdots & \vdots & \vdots \\ 1 & 1 & X_n \end{bmatrix}$$

As a result, the model becomes:

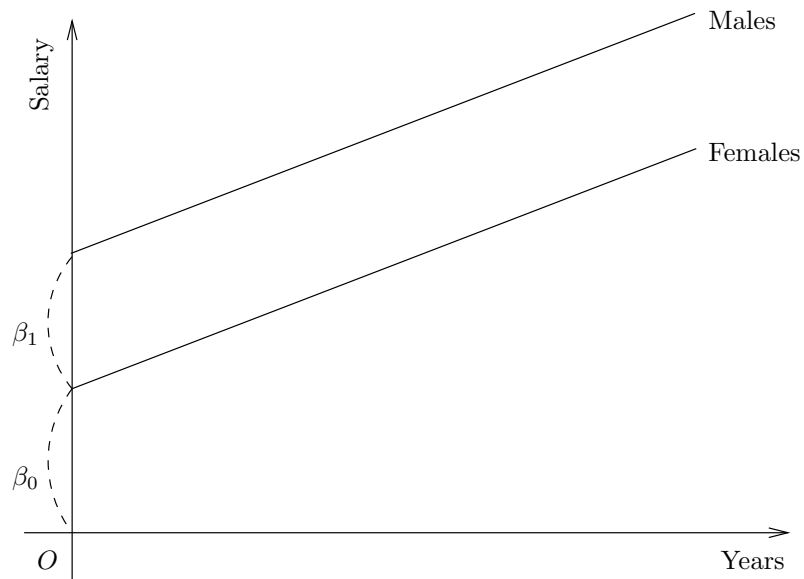
$$Y_i = \begin{cases} (\beta_0 + \beta_1) + \beta_2 X_i + \epsilon_i & \text{ male} \\ \beta_0 + \beta_2 X_i + \epsilon_i & \text{ female} \end{cases}.$$

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

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

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.

Categorical and continuous predictors



Effect of years is same for both genders.

Effect of gender is same for all years.

β_0 = intercept for females

β_1 = intercept for males – intercept for females

= gender effect at any years

β_2 = slope for both

3. Polynomial regression

A polynomial regression model with one predictor variable

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \quad (6.2)$$

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:

$$Y = \beta_0 + \beta_1 X_1^{\beta_3} + \beta_2 X_2^{\beta_4} + \epsilon$$

$$Y = \frac{\beta_0}{1 + e^{-(\beta_1 X_1 + \beta_2 X_2)}} + \epsilon.$$

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, \dots, 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

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

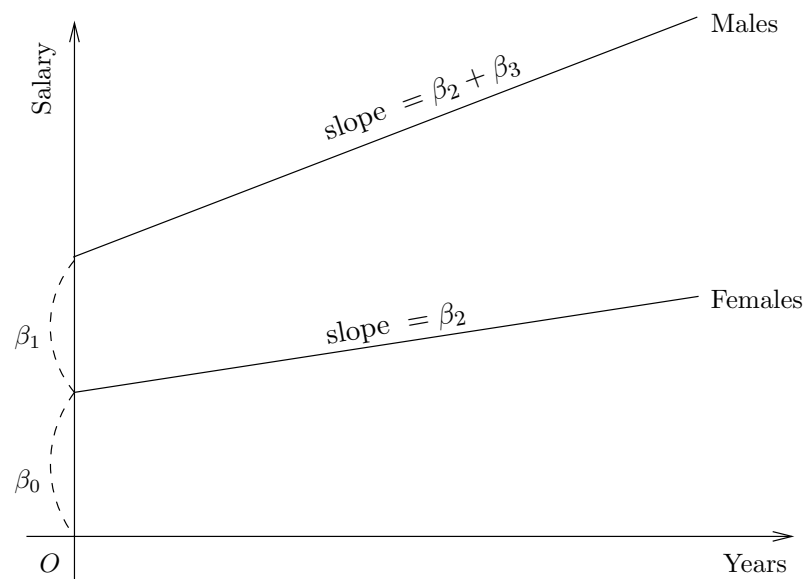
Let Z be defined as follows:

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

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



β_0 = intercept for females

β_1 = intercept for males – intercept for females

= gender effect “at years = 0”

β_2 = slope for females

= years effect for females

β_3 = slope for males – slope for females

= years effect for females – 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.

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:

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

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

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

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 $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$.

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

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

and it has a normal distribution

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

4 Estimation of regression parameters

1. Least-squares method

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

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

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

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

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

2. Using the projection

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

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

3. MLE

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

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

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

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

These estimators are least-squares and MLE.

5 Fitted values and residuals

The fitted values \hat{Y}_i are

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

In matrix notation, we have

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

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

$$\hat{\mathbf{Y}} = \mathbf{X} \overbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}}^{\mathbf{H}} \underbrace{\hspace{1.5cm}}_{\hat{\boldsymbol{\beta}}}$$

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

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

Then we have

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

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

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

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

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

and is estimated by

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

where

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

6 ANOVA results

6.1 ANOVA table

The sums of squares can be expressed in matrix notation.

$$\begin{aligned} \text{SSTo} &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = \|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2 \\ &= \sum_{i=1}^n Y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n Y_i \right)^2 = \mathbf{Y}'\mathbf{Y} - \frac{1}{n} \mathbf{Y}'\mathbf{J}\mathbf{Y} \\ \text{SSE} &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 \\ &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} \\ \text{SSR} &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \|\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}\|^2 \\ &= \text{SSTo} - \text{SSE} \\ &= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - \frac{1}{n} \mathbf{Y}'\mathbf{J}\mathbf{Y} \end{aligned}$$

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

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

$$\text{SSTo} = \text{SSE} + \text{SSR}$$

ANOVA table

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

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:

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

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

The above test is equivalent to testing

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

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

The decision rule at the significance level α is :

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

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

Remark 6.2. Notes about overall F -test:

1. If H_0 is rejected, we know that at least one of the coefficients $\beta_1, \dots, \beta_{p-1}$ 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 $\beta_1, \dots, \beta_{p-1}$ 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

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

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.

△

6.3 Coefficient of multiple determination

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

$$R^2 = \frac{\text{SSR}}{\text{SST}_0} = 1 - \frac{\text{SSE}}{\text{SST}_0}$$

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

Remark 6.3. 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 $p = 2$ (intercept and one predictor), 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:

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n - p)}{\text{SSTo}/(n - 1)} = 1 - \left(\frac{n - 1}{n - p}\right) \frac{\text{SSE}}{\text{SSTo}} = 1 - \left(\frac{n - 1}{n - p}\right)(1 - R^2).$$

△

7 Inferences in regression analysis

1. Regression coefficient.

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

$$\begin{aligned}\text{Cov}(\hat{\beta}) &= \text{Cov}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Cov}(\mathbf{Y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

$$\widehat{\text{Cov}}(\hat{\beta}) = \text{MSE}(\mathbf{X}'\mathbf{X})^{-1}$$

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

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

2. Interval estimation of β_j .

For the normal error regression model, we have

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

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

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

3. Test for β_j .

To test

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

we can use the test statistic:

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

and the decision rule:

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

Otherwise conclude H_1 .

4. *Bonferroni joint confidence intervals.*

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

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

5. *Mean response at X_h .*

Define $\mathbf{x}'_h = [1 \ X_{h1} \ X_{h2} \ \dots \ X_{h,p-1}]$. Then the fitted value at $X_{h1}, X_{h2}, \dots, X_{h,p-1}$ in matrix notation is

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

Hence we have

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

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

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

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

Thus, the above variance in matrix notation becomes

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

Hence, we have

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

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.

Minitab with SSE subcommand

```
1 MTB> set c5
2 MTB> -10:10
3 MTB> end
4 MTB> let c5 = c5/10
5
6 MTB> %S:\LM\BOXCox c4 c1-c3 c5 c6;
7 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.

Plot of SSE

```
1 > # Data generation
2 > x1 = 1:5
3 > x2 = sqrt(1:5)
4 > x3 = (1:5)^2
5 > y = exp( x1 + 2*x2 + 3*x3)
6 >
7 > # SSE values for each of lambda
8 > lam = seq(-0.005, 0.005, 0.0001)
9 > SSE = BoxCox ( y ~ x1 + x2 + x3, lambda=lam, SSE=TRUE )
10 > plot(lam, SSE, type="l")
```

9 Example: Patient-satisfaction data

We shall develop a multiple regression application with three predictors. We will analyze the Problems 6.15 ~ 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.*

Read Data

```
1 MTB > read c4 c1-c3;
2 SUBC>      file "S:\LM\CH06PR15.TXT" .
3 Entering data from file: S:\LM\CH06PR15.TXT
4 46 rows read.
5 MTB > NAME c1 "Age"
6 MTB > NAME c2 "Severity"
7 MTB > NAME c3 "Anxiety"
8 MTB > NAME c4 "Satisfaction"
```

2. *Scatter plot matrix.*

Scatter plot

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

3. *Correlation matrix.*

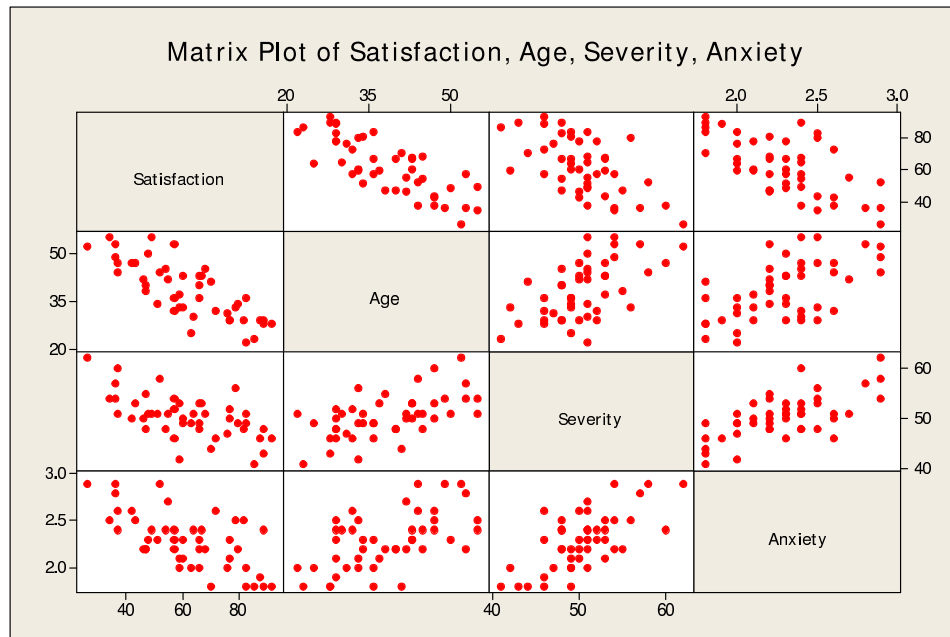
Correlation matrix

```
1 MTB > correlation c4 c1-c3
2 Correlations: Satisfaction, Age, Severity, Anxiety
3
4           Satisfaction           Age           Severity
5 Age           -0.787
6               0.000
7
8 Severity       -0.603           0.568
9               0.000           0.000
10
11 Anxiety        -0.645           0.570           0.671
12               0.000           0.000           0.000
13
14 Cell Contents: Pearson correlation
15                P-Value
```

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

Regression

```
1 MTB > regr c4 3 c1 c2 c3;
2 SUBC> fits c5;
```



```

3 SUBC> resid c6.
4
5 Regression Analysis: Satisfaction versus Age, Severity, Anxiety
6
7 The regression equation is
8 Satisfaction = 158 - 1.14 Age - 0.442 Severity - 13.5 Anxiety
9
10 Predictor      Coef    SE Coef      T      P
11 Constant      158.49    18.13     8.74   0.000
12 Age           -1.1416    0.2148    -5.31   0.000
13 Severity      -0.4420    0.4920    -0.90   0.374
14 Anxiety       -13.470    7.100    -1.90   0.065
15
16 S = 10.0580    R-Sq = 68.2%    R-Sq(adj) = 65.9%
17
18 Analysis of Variance
19 Source         DF         SS         MS         F         P
20 Regression      3      9120.5      3040.2     30.05   0.000
21 Residual Error  42      4248.8       101.2
22 Total          45     13369.3
23
24 Source         DF      Seq SS
25 Age            1     8275.4
26 Severity       1      480.9
27 Anxiety        1      364.2

```

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

Confidence interval

```

1 MTB > regress c4 3 c1-c3;
2 SUBC> predict 35 45 2.2;
3 SUBC> confidence 90.
4

```

5	Predicted Values for New Observations				
6					
7	New				
8	Obs	Fit	SE Fit	90% CI	90% PI
9	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, \dots, X_{p-1} in the following way:

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

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

$$X_{BP}^2 = \frac{SSR^*}{2} \div \left(\frac{SSE}{n} \right)^2 \dot{\sim} \chi_{df=p-1}^2,$$

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

Regression of $\hat{\epsilon}_i^2$ on X_1, \dots, X_{p-1}

```

1 MTB > let c22=c6*c6
2 MTB > regr c22 3 c1-c3
3
4 Regression Analysis: C22 versus Age, Severity, Anxiety
5
6 The regression equation is
7 C22 = 49 - 2.73 Age + 3.81 Severity - 19.4 Anxiety
8
9 Predictor      Coef    SE Coef      T      P
10 Constant      49.4     167.5     0.30   0.769
11 Age           -2.728    1.984    -1.37   0.176
12 Severity       3.807    4.545     0.84   0.407
13 Anxiety       -19.38    65.59    -0.30   0.769
14
15 S = 92.9193    R-Sq = 5.6%    R-Sq(adj) = 0.0%
16
17 Analysis of Variance
18 Source         DF      SS      MS      F      P
19 Regression       3    21356    7119    0.82   0.488
20 Residual Error  42   362628   8634
21 Total           45   383983
22
23 Source      DF  Seq SS

```

24	Age	1	15085
25	Severity	1	5517
26	Anxiety	1	754

Remark 6.4. From the Minitab results, we have the test statistic

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

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

Comparing $X_{BP}^2 = 1.25$ 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>

BPtest.MAC

```

1 MTB > read c4 c1-c3;
2 SUBC> file "S:\LM\CH06PR15.TXT" .
3 Entering data from file: S:\LM\CH06PR15.TXT
4 46 rows read.
5
6 MTB > %S:\LM\BPtest c4 c1-c3 .
7 Executing from file: S:\LM\BPtest.MAC
8
9 Data Display
10
11 Breusch-Pagan Test Statistic:      1.25157
12 Degrees of Freedom:              3
13 p-value:                          0.74066

```

R

1. *Read the data.*

Read Data

```

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

```

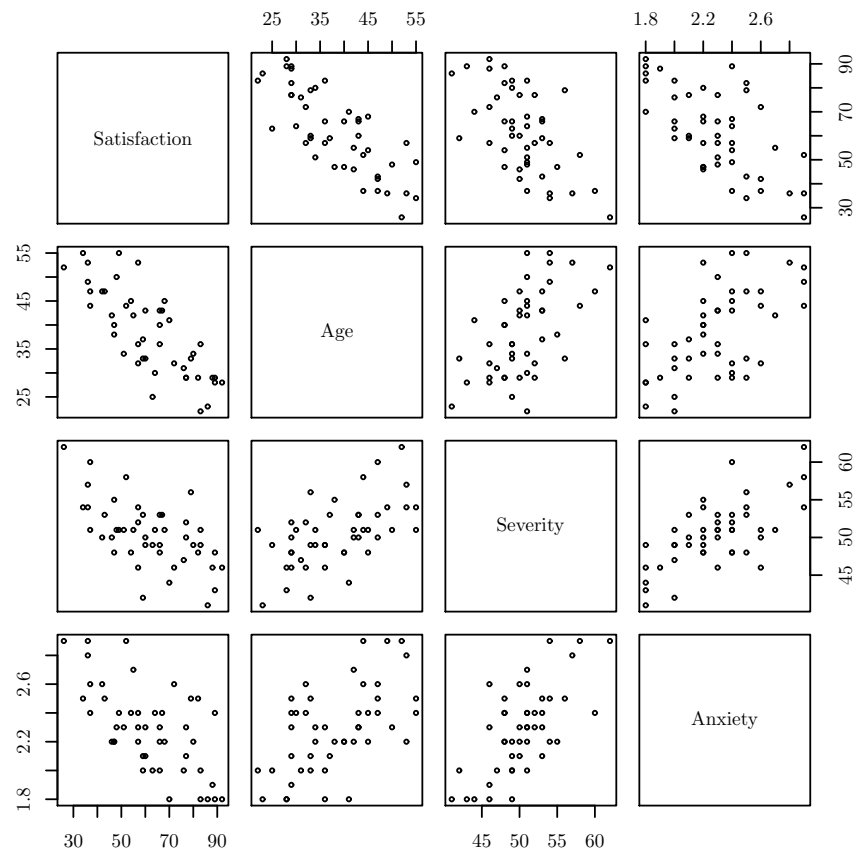
2. *Scatter plot matrix.*

Scatter plot

```

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

```



3. Correlation matrix.

Correlation matrix

```

1 > cor( mydata )
2           Satisfaction      Age      Severity      Anxiety
3 Satisfaction      1.0000000 -0.7867555 -0.6029417 -0.6445910
4 Age               -0.7867555  1.0000000  0.5679505  0.5696775
5 Severity          -0.6029417  0.5679505  1.0000000  0.6705287
6 Anxiety           -0.6445910  0.5696775  0.6705287  1.0000000

```

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

Regression

```

1 > LM = lm ( y ~ x1 + x2 + x3 )
2 > summary(LM)
3 Call:
4 lm(formula = y ~ x1 + x2 + x3)
5
6 Residuals:
7      Min       1Q   Median       3Q      Max
8 -18.3524  -6.4230   0.5196   8.3715  17.1601
9
10 Coefficients:
11             Estimate Std. Error t value Pr(>|t|)
12 (Intercept)  158.4913    18.1259   8.744 5.26e-11 ***
13 x1           -1.1416     0.2148  -5.315 3.81e-06 ***
14 x2           -0.4420     0.4920  -0.898  0.3741

```

```

15 x3          -13.4702      7.0997  -1.897   0.0647 .
16 ---
17 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
18
19 Residual standard error: 10.06 on 42 degrees of freedom
20 Multiple R-Squared:  0.6822,    Adjusted R-squared:  0.6595
21 F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
22
23 > anova (LM)
24 Analysis of Variance Table
25 Response: y
26      Df Sum Sq Mean Sq F value    Pr(>F)
27 x1      1 8275.4   8275.4  81.8026 2.059e-11 ***
28 x2      1  480.9    480.9   4.7539  0.03489 *
29 x3      1   364.2    364.2   3.5997  0.06468 .
30 Residuals 42 4248.8    101.2
31 ---
32 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

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

Confidence interval

```

1 > new = data.frame( x1=35, x2=45, x3=2.2 )
2 > predict ( LM, newdata=new, interval="confidence", level=0.90)
3           fit          lwr          upr
4 [1,] 69.01029 64.52854 73.49204
5 >
6 > new = data.frame( x1=35, x2=45, x3=2.2 )
7 > predict ( LM, newdata=new, interval="prediction", level=0.90)
8           fit          lwr          upr
9 [1,] 69.01029 51.50965 86.51092

```

6. Tests for constancy of error variance.

We present the Breusch-Pagan test using R.

Breusch-Pagan

```

1 > e = resid(LM)
2 > SSE = sum( e^2 )
3 > sigma2 = e^2
4 >
5 > LM2 = lm ( sigma2 ~ x1 + x2 + x3 )
6 > SSR.star = sum( (fitted(LM2)-mean(sigma2))^2 )
7 >
8 > n = length(y)
9 >
10 > cbind(SSR.star, SSE, n)
11      SSR.star      SSE      n
12 [1,] 21355.53 4248.841 46
13 >
14 > X.BP = SSR.star/2 / ( (SSE/n)^2 )
15 >
16 > X.BP
17 [1] 1.251570
18 >
19 > qchisq(0.99, df = 3) ## chi-square critical value
20 [1] 11.34487

```

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

<https://github.com/AppliedStat/LM>

Breusch-Pagan.R

```
1 > # 1. Read the data
2 > url = "https://raw.githubusercontent.com/AppliedStat/LM/master/CH06PR15.txt"
3 > mydata=read.table(url)
4
5 > y = mydata[,1]
6 > x1 = mydata[,2]
7 > x2 = mydata[,3]
8 > x3 = mydata[,4]
9
10 > # 2. Read Breusch-Pagan test R program
11 >
12 >     source("https://raw.githubusercontent.com/AppliedStat/LM/master/Breusch-Pagan.R")
13 > BP.test ( y ~ x1 + x2 + x3 )
14 $test.stat
15 [1] 1.251570
16
17 $df
18 [1] 3
19
20 $p.value
21 [1] 0.7406642
```
