MULTIPLE REGRESSION AND THE GENERAL LINEAR MODEL

We extend the simple linear regression case to the multiple linear regression case to handle situations where the dependent variable y is assumed to be related to more than one independent variable, say, $x_1, x_2, \ldots, x_k$.

The model we use is called a `general linear model` – its form is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

Our assumptions about the ϵ's are the same, namely the ϵ's are independent with mean $E(\epsilon)$ zero and variance σ_ϵ^2, and they are normally distributed. Data consist of n `cases` of $k+1$ values denoted by

$$(y_i,\, x_{i1},\, x_{i2}, \ldots, x_{ik}), \quad i = 1, 2, \ldots, n$$

Using this full notation we can write the model as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

The form of the general linear model, as given, is understood to permit polynomial terms like x_2^3, cross-product terms like $x_2 x_3$ – etc. – Thus the polynomial model in one variable

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

fits the form of the general linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

with $x_1 \equiv x,\ x_2 \equiv x^2,\ x_3 \equiv x^3$.

The multiple regression model

$$\begin{aligned} y \;\; = \;\; & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 \\ & + \beta_5 x_1^3 + \beta_6 x_2^3 + \beta_7 x_1 x_2 + \epsilon \end{aligned}$$

can also be written as

$$\begin{aligned} y \;\; = \;\; & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \\ & + \beta_6 x_6 + \beta_7, x_7 + \epsilon \end{aligned}$$

with $x_3 = x_1^2,\ x_4 = x_2^2,\ x_5 = x_1^3,\ x_6 = x_2^3,\ x_7 = x_1\,x_2$.

Even terms like $\log(x_3)$, e^{x_2}, $\cos(x_4)$, etc., are permitted.

The model is called a `linear model` because its expression is linear in the β's. The fact that it may involve nonlinear functions of the x's is irrelevant.

The concept of `interaction` will be discussed later. Here let us just state that the presence (or absence) of interaction between two independent variables (say) x_1 and x_2 can be tested by including product terms such as $x_1\,x_2$ in the model. Thus the model:

$$y = \beta_0 + \beta_1\,x_1 + \beta_2\,x_2 + \beta_3\,x_1\,x_2$$

allows us to test whether interaction between x_1 and x_2 exists. The interpretation of this model is postponed till later.

When a general linear model relates y to a set of `quantitative` independent variables (x's) that may include squared terms, product terms, etc., we have a `multiple regression model`. When the independent variables are `dummy variables coded to 0 or 1` representing values of `qualitative` independent variables or `levels` of treatment factors, the resulting models are called `analysis of variance` models.

Just as we did in Chapter 11 for the simple linear regression model $y = \beta_0 + \beta_1\,x + \epsilon$, we will consider least squares estimation of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ in the general linear model. These will be denoted as $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

Then confidence intervals and tests of hypotheses concerning the parameters will be discussed as well as inferences about $E(y)$ and a future y observation.

Computing quantities of interest like the $\hat{\beta}_i$, SSE, SSREG, MSE, MSREG, -etc- must be done using a computer because the amount of arithmetic required is very large. In order to show you exactly what computations must be done (so you will understand what the computer is doing) it is easiest to use matrix notation.

Multiple Regression Computations Using Matrix Notation

Using matrix notation, we can define vectors $\boldsymbol{y}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ and the matrix $\boldsymbol{X}$, as follows with the objective being to give a compact form of definition of quantities like least squares point and interval estimates, test statistics, etc.

These are defined as follows:

$$\underset{n\times 1}{\boldsymbol{y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \underset{n\times(k+1)}{\boldsymbol{X}} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ 1 & x_{31} & x_{32} & \dots & x_{3k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

$$\underset{(k+1)\times 1}{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \underset{n\times 1}{\boldsymbol{\epsilon}} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Now the model can be written in terms of matrices as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The following quantities, defined in terms of the above, are useful in multiple regression computations that follow:

$\boldsymbol{X}'\boldsymbol{X}$ is the matrix of sums of cross products of the various columns of $\boldsymbol{X}$. The element in the i^{th} row and the j^{th} column of $\boldsymbol{X}'\boldsymbol{X}$ is shown as

$$\boldsymbol{X}'\boldsymbol{X} = \left(\sum_{t=1}^{n} x_{ti}\, x_{tj}\right)$$

Denote the elements of the inverse of this matrix by v_{ij} where $i = 1, 2, \dots, k,\ j = 1, 2, \dots, k$. In this notation

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} v_{00} & v_{01} & v_{02} & \dots & v_{0k} \\ v_{10} & v_{11} & v_{12} & \dots & v_{1k} \\ v_{20} & v_{21} & v_{22} & \dots & v_{2k} \\ \vdots & & & & \\ v_{k0} & v_{k1} & v_{k2} & \dots & v_{kk} \end{bmatrix}$$

(Note: $v_{ij} = v_{ji}$ for all (i, j) i.e., the $(X'X)^{-1}$ matrix is symmetric.)

The least squares estimates $\hat{\beta}_i$ of the parameters β_i in the model are elements of the vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$ where $\hat{\boldsymbol{\beta}}$ is the solution to the `normal equations`

$$\boldsymbol{X}'\boldsymbol{X}\,\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$$

and is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

The standard deviations of the $\hat{\beta}_i$ are

$$\begin{aligned} \sigma_{\hat{\beta}_0} &= \sigma_\epsilon \sqrt{v_{00}} \\ \sigma_{\hat{\beta}_1} &= \sigma_\epsilon \sqrt{v_{11}} \\ &\vdots \\ \sigma_{\hat{\beta}_k} &= \sigma_\epsilon \sqrt{v_{kk}} \end{aligned}$$

The sum of squares of residuals (or SSE) is

$$\text{SSE} = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$$

and the mean square for error, i.e., the s_ϵ^2 which we use to estimate σ_ϵ^2 is

$$s_\epsilon^2 = \text{MSE} = \frac{\text{SSE}}{\text{n} - (\text{k}+1)} \ .$$

Thus the standard errors of $\hat{\beta}_j$ are:

$$\begin{aligned} s_{\hat{\beta}_0} &= s_\epsilon \sqrt{v_{00}} \\ s_{\hat{\beta}_1} &= s_\epsilon \sqrt{v_{11}} \\ &\vdots \\ s_{\hat{\beta}_k} &= s_\epsilon \sqrt{v_{kk}} \end{aligned}$$

Confidence intervals for the β_j therefore have the form

$$\hat{\beta}_j \pm t_{\alpha/2} \cdot s_\epsilon \sqrt{v_{jj}}\,, \quad j = 0, 1, 2, \ldots, k$$

where $t_{\alpha/2}$ is the upper $(1 - \alpha/2)$ percentile of the t-distribution with $df = n - (k+1)$.

To test $H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$ use the t-statistic

$$t_j = \frac{\hat{\beta}_j}{s_\epsilon \sqrt{v_{jj}}}$$

with $df = n - (k+1)$, for each $j = 1, \ldots, k$.

A 100(1-α)% confidence interval for $E(y_{n+1})$, the population mean of Y_{n+1} at $(x_{n+1,1},\ x_{n+1,2}, \ldots, x_{n+1,k})$

$$\hat{y}_{n+1} \pm t_{\alpha/2} \cdot s_\epsilon \sqrt{\boldsymbol{\ell}'_{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\ell}_{n+1}}$$

where $\boldsymbol{\ell}'_{n+1} = (1,\, x_{n+1,1},\, x_{n+1,2}, \ldots, x_{n+1,k})$

A 100(1-α)% prediction interval for a new y_{n+1} simply adds one under the square root above.

$$\hat{y_{n+1i}} \pm t_{\alpha/2}\, s_\epsilon \sqrt{1 + \boldsymbol{\ell}'_{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\ell}_{n+1}}$$

The `coefficient of determination` $R^2_{y\cdot x_1,x_2,\ldots,x_k}$ is computed as

$$R^2_{y\cdot x_1,x_2,\ldots,x_k} = \frac{S_{yy} - \text{SSE}}{S_{yy}}$$

Generally, $R^2_{y\cdot x_1,\ldots,x_k}$ is the proportion of the variability in y that is accounted for by the independent variables $x_1, x_2, \ldots, x_k$ in the model. Without using matrix notation the `normal equations` are the following

$$\begin{array}{llllll}
\sum y = & n\,\hat{\beta}_0 & +\sum x_1\,\hat{\beta}_1 & +\cdots+ & \sum x_k\,\hat{\beta}_k \\
\sum x_1\,y = & \sum x_1\,\hat{\beta}_0 & +\sum x_1^2\,\hat{\beta}_1 & +\cdots+ & \sum x_1\,x_k\,\hat{\beta}_k \\
\vdots & \vdots & \vdots & \cdots & \vdots \\
\sum x_k\,y = & \sum x_k\,\hat{\beta}_0 & +\sum x_k\,x_1\,\hat{\beta}_1 & +\cdots+ & \sum x_k^2\,\hat{\beta}_k
\end{array}$$

More Inference On Multiple Regression Model

1. In the general linear model the β's are sometimes called `partial slopes`. β_j is the expected change in y for unit change in x_j `when all other` x`'s are held constant` i.e., when the the other variables are fixed at a set of values.
2. `Analysis of variance table for multiple regression`

Source	df	Sum of Squares	Mean Square	F
Regression	k	SSReg	MSReg=SSReg/k	F=MSReg/MSE
Error	n-(k+1)	SSE	MSE=SSE/(n-(k+1))	
Total	n-1	SSTot		

3. Quantities like $s_{\hat{\beta}_j}$, and confidence interval (band) for $E(y)$, or prediction interval (band) for a future y cannot reasonably be computed by hand. To obtain intervals at x values for which y was not observed, include an $(n+1)^{th}$ row of x values in the data set with a missing value (i.e. a period) for the value of y. That is the line of data is in the form

$$(\cdot,\, x_1, x_2, \ldots, x_k)$$

(Note that when we say "band" for $E(y)$ and future y we are talking about a multivariate band, i.e., a "band" in more than two dimensions.)

Using the period in the y-value position tells SAS the y observation is missing. SAS will ignore the row of $(k+1)$ x values when computing sums of squares, least squares estimates, etc., but will the compute predicted value $\hat{y}$, confidence interval for $E(y)$, and prediction interval for future y observation at the given set of $(k+1)$ x values.

4. Other than a `full model` (i.e.,one involving all the x variables being considered at the time) one often looks at `reduced models` (ones involving some but not all of the full set of independent variables.) `Estimates computed for full and reduced models will usually all be different`.

   If the x_1 variable is in both models, for example, the estimate of the β_1 coefficient ($\hat{\beta}_1$) will not be the same for the two least squares fits. The Example 12.18 (page 658) illustrates this.

5. The term `multicollinearity` is used to describe the situation wherein some of the independent variables have strong correlations amongst themselves.

   If we are using both the high school GPA (x_1), and the SAT score (x_2) in a multiple regression model to predict college GPA (y) , it is very likely that the two variables x_1 and x_2 are correlated.

   If correlation is very high for a pair of x variables the effect is that the estimated coefficients ($\hat{\beta}$'s) of one or both of these variables will have very high standard errors, resulting in the t-tests for these variables $H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$ failing to reject the respective null hypotheses.

   The investigator may make a decision on which of these variables will be retained in the model. Such a decision is usually based on subject matter knowledge about the predictive value of these variables to the model. Variables may be retained in the model even in the presence of some multicollinearity if they contribute types of information needed to explain the variability in y other variables in the model may not be able to provide.

Other Formulas Useful for Interpretation

Your textbook gives the estimated standard errors of the $\hat{\beta}_j$'s as functions of R^2 values obtained by regressing x_j on the rest of the x variables.

$$s_{\hat{\beta}_j} = s_\epsilon \sqrt{\frac{1}{S_{x_j x_j}(1 - R^2_{x_j \cdot x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k})}}$$

where

$$S_{x_j x_j} = \sum x_j^2 - \frac{(\sum x_j)^2}{n}$$

and

$$R^2_{x_j \cdot x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k}$$

is the coefficient of determination computed using the variable x_j as dependent variable along with independent variable $x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ (i.e, all the other x's in the original model.)

This definition is useful for illustrating the effect of `multicollinearity` on the estimation of each coefficient β_j. If the x_j variable is highly `collinear` with one or more of the other x variables, $R^2_{x_j \cdot x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k}$) will be large. Thus the denominator of the formula above for $s_{\hat{\beta}_j}$ will be small, resulting in possibly a very large standard error for $\hat{\beta}_j$ showing that the estimate $\hat{\beta}_j$ will be inaccurate. Thus this formula shows directly the effect of severe multicollinearity on the estimation of the coefficients.

The quantity `variance inflation factor` for a coefficient β_j is defined as

$$\text{vif}_j = \frac{1}{1 - R^2_{x_j \cdot x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k}}$$

and measures the factor by which the variance (square of the standard error) of $\hat{\beta}_j$ increases because of multicollinearity. If the `vif` is 1, there is no multicollinearity problem at all. As a rule of thumb a `vif` of greater than 10 indicates severe collinearity of the x_j with other x's and the standard error of the corresponding estimated coefficient will be inflated.

One noticeable effect of high standard error of $\hat{\beta}_j$ is that it may result in an incorrect sign for $\hat{\beta}_j$. For example, They are one would expect the sign of $\hat{\beta}_1$ to be positive if subject matter knowledge suggests that the mean $E(Y)$ must increase when the value of x_1 increases (when the other x variable values are unchanged); however, if the parameter is estimated with high variabilty, it may be estimated as a negative value.

Inferences Concerning a Set of β's

We know that we can test hypothesis about any one of the β's in the model using a t-test. i.e., a hypothesis like $H_0 : \beta_2 = 0$ employs the test statistic $t = \hat{\beta}_2 / s_{\hat{\beta}_2}$.

Sometimes we wish to test a hypothesis whether two or more of the β's in the model are zero against the alternative hypothesis that at least one of these β's is not zero. Here we shall assume that k is the total number of predictors (x's) and that g is the number of coefficients considered to be `nonzero` where, obviously, $g < k$. Thus $k - g$ represents the number of variables with coefficients hypothesized to be zero. To test this hypothesis we use an F test. To formulate it we first fit the two models:

`Full Model:`

The model involving all x's.

`Reduced Model:`

The model involving only those x's from the full model whose β coefficients are not hypothesized to be zero.

It is easier to state the hypotheses needed to be tested if the $k - g$ variables corresponding to the coefficients to be tested equal to zero are arranged to appear last in the full model. In this case, the hypotheses can be stated as

$H_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$

H_a : at least one of these is not zero.

Now the F-test statistic for testing H_0 is

$$F = \frac{[\text{SSReg(Full)} - \text{SSReg(Reduced)}]/(\text{k} - \text{g})}{\text{MSE(Full)}}$$

The numerator and denominator degrees of freedom here for the F statistic are $k - g$ and $n - (k + 1)$, respectively. When this computed F exceeds the F-value obtained from the F-tables for a chosen α, H_0 is rejected.

Example 12.18 illustrates this kind of a test. The two models considered are:

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$

- $y = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \epsilon$

The null hypothesis $H_0 : \beta_3 = \beta_4 = 0$ tests whether x_3 and x_4 taken together contribute significant predictive value to be included in the Full Model . Fitting the two models, using the same 20 observations in each case, yields the following:

Full Model:

Prediction Equation:

$$\hat{y} = -2.784 + 0.02679x_1 + 0.50351x_2 + 0.74293x_3 + 0.05113x_4$$

ANOVA Table:

Source	df	SS of	MS	F
Regression	4	24.0624	6.0156	39.65
Error	15	2.27564	0.1517	
Total	19	26.338		

Reduced Model:

Prediction Equation:

$$\hat{y} = -0.8709 + 0.0394x_1 + 0.828x_2$$

ANOVA Table:

Source	df	SS of	MS	F
Regression	2	2.9125	1.456	1.06
Error	17	23.425	1.378	
Total	19	26.338		

Thus the test for $H_0 : \beta_3 = \beta_4 = 0$ is:

$$\begin{aligned} \text{T.S.} \qquad F &= \frac{[\text{SSReg(Full)} - \text{SSReg(Reduced)}]/(\text{k} - \text{g})}{\text{MSE(Full)}} \\ &= \frac{(24.0624 - 2.9125)/(4 - 2)}{.1517} = 69.7 \end{aligned}$$

For $\alpha = 0.01$, the percentile from the F table is $F_{.01,2,15} = 6.36$. Thus we reject H_0 at $\alpha = 0.01$. In this case we have evidence that Access and Structure variables do actually contribute predictive value to the model.

This test is applicable to many other kinds of situations. Assessing the viability of higher-order terms and their interactions is another possible application of a test like this.

Note that SSReg(Full) can never be less than SSReg(Reduced) because if it was, then the reduced model would explain more variability in y than the full model. This is not possible since the full model includes the reduced model plus additional x variables.

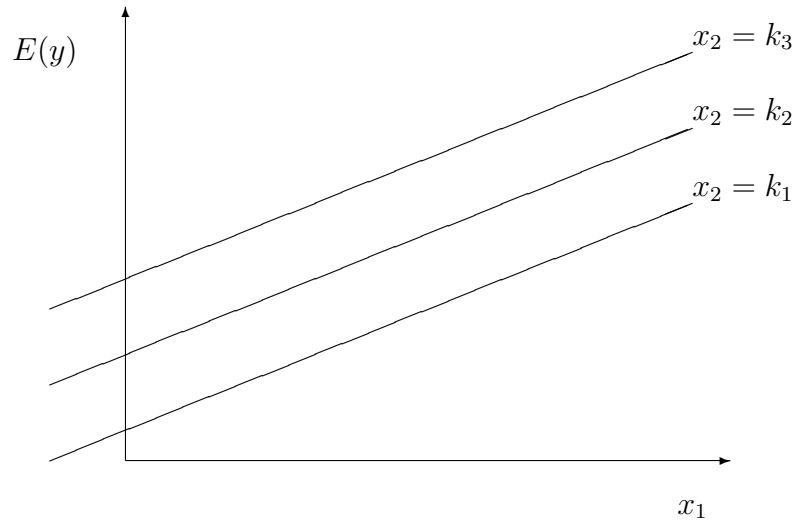
Including Interaction Terms in the Model

The concept of `interaction` between two independent variables (say) x_1 and x_2 is discussed in Section 12.1. By definition x_1, and x_2 `do not interact` if the expected change in y for unit change in x_1, does not depend on x_2.

To illustrate, consider the first-order model in only x_1 and x_2,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

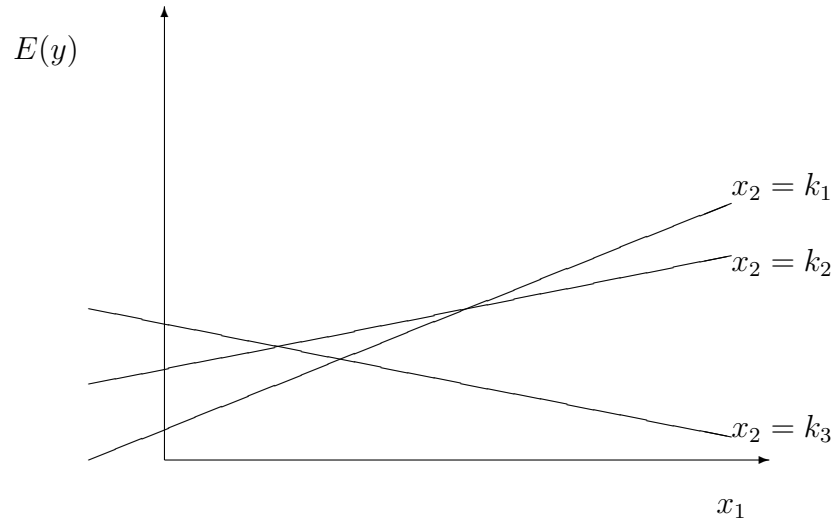
i.e., consider $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$



When the value of x_2 is kept fixed, $E(y)$ is a straight line function of x_1. Different values of x_2 give parallel lines. Thus the expected change in y for unit change in x_1 will remain β_1 regardless of what the value of x_2 is.

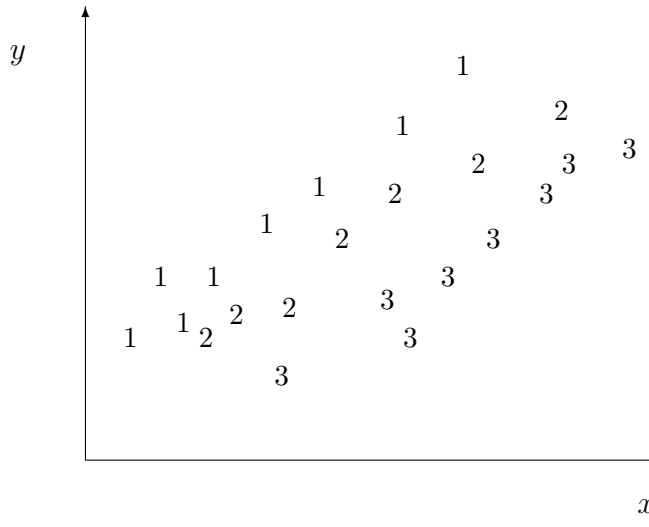
Now suppose we have the model

$$\begin{aligned} E(y) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \\ &= \beta_0 + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2 \end{aligned}$$

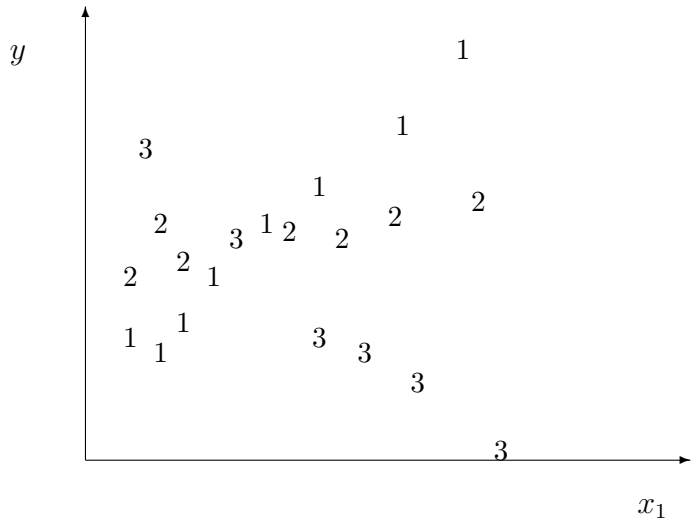


When the value of x_2 is kept fixed, $E(y)$ is still a straight line function of x_1. However, the slope is $\beta_1 + \beta_3 x_2$, which obviously is a function of x_2. Therefore the slope now depends on the value of x_2. So, as you see in the picture, the lines are not parallel as the slopes are different for different values of x_2. So x_1 and x_2 are said to interact.

To detect interaction in regression data, we can construct a scatterplot of the data with the numbers identifying values of the variable x_2:



Here the scatterplot suggests parallel traces, hence no interaction is indicated. But if the plot looks like the one below interaction is indicated.



A term in the model like $\beta x_2 x_3$ is used to account for interaction (a 2-factor interaction term, we say) between x_2 and x_3. Higher order interactions (3-factor, 4-factor, etc.) are conceivable, like $\beta x_1 x_3 x_4$ or $\beta x_1 x_2 x_3 x_4$, but we seldom consider other than possibly 2-factor interactions, mainly because such models become difficult to interpret and therefore not useful.

Comparing Slopes of Regression Lines

When we want to use the general linear model to test whether the slopes of regression lines under different conditions (or levels of treatments) we may include a quantitative independent variable (i.e., an x variable) of a special type called a `dummy` variable (or an `indicator` variable) in the model. This variable takes values $0, 1, ..etc.$ to "indicate" the different conditions or different levels of a treatment.

Example 12.21, 12.22

An anxiety score y is taken from rats given two drug products (A and B) with 10 rats assigned to each of three doses $(5, 10,$ and 20mg)of each drug. The data are:

TABLE 12.4
Rat anxiety scores

Drug Product	Drug Dose (mg)		
	5	10	20
A	15 16	18 16	20 17
	16 15	17 15	19 18
	18 16	18 19	21 21
	13 17	19 18	18 20
	19 15	20 16	19 17
	av = 16	av = 17.6	av = 19.0
B	16 15	19 18	24 23
	17 15	21 20	25 24
	18 18	22 21	23 22
	17 17	23 22	25 26
	15 16	20 19	25 24
	av = 16.4	av = 20.5	av = 24.1

Let y denote the response, anxiety score and x_1 denote the drug dose, which is a quantitative variable. However, we require a qualitative variable (say, x_2) to represent the drug product since it has values A and B.

We set $x_2 = 0$ if a response is from a rat administered drug A, and $x_2 = 1$ if it is from drug B. Thus the actual data looks as follows:

y	x1	x2
15	5	0
15	5	0
:	:	:
18	10	0
:	:	:
16	5	1
17	5	1
:	:	:
19	10	1
:	:	:
24	20	1

The model includes an interaction term that, as we see below, will allow us to test whether the two regression lines corresponding to the two drug products have the same slope:

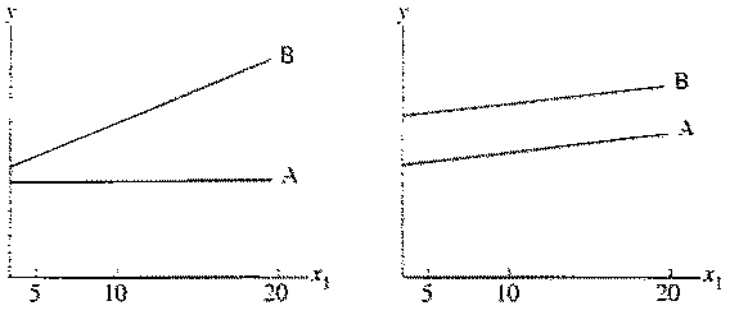
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

Thus the expected value (or mean) response $E(Y)$ for each drug is obtained by substituting $x_2 = 0$ and $x_2 = 1$, respectively, in the above equation:

$$\begin{aligned} \text{drug A:} \quad E(Y) &= \beta_0 + \beta_1 x_1 \\ \text{drug B:} \quad E(Y) &= \beta_0 + \beta_1 x_1 + \beta_2 + \beta_3 x_1 \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 \end{aligned}$$

These two equations are linear regression lines, each corresponding to the set of data for the two drugs. If β_3 in the second was zero then the two lines would have the same slope, but different y-intercepts. The situation is illustrated in the figure below:

FIGURE 12.4
Comparing two regression lines



(a) $\beta_3 \neq 0$: interaction is present: intersecting lines

(b) $\beta_3 = 0$: interaction is not present: parallel lines

We can test the hypothesis $H_0 : \beta_3 = 0$ by comparing the `full model`

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

with the `reduced model:`

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

as in the previous section. Taking values from the SAS output, the test for $H_0 : \beta_3 = 0$ is:

$$\begin{aligned} \text{T.S.} \quad F &= \frac{[\text{SSReg(Full)} - \text{SSReg(Reduced)}]/(\text{k} - \text{g})}{\text{MSE(Full)}} \\ &= \frac{(442.10 - 389.60)/(3 - 2)}{2.38622} = 22.00 \end{aligned}$$

We reject $H_0 : \beta_3 = 0$ at $\alpha = 0.05$ as the computed F exceeds $F_{.05,1,56} \approx 4.0$ and conclude that the slopes of the two regression lines are different.

The interpretation of significant interaction

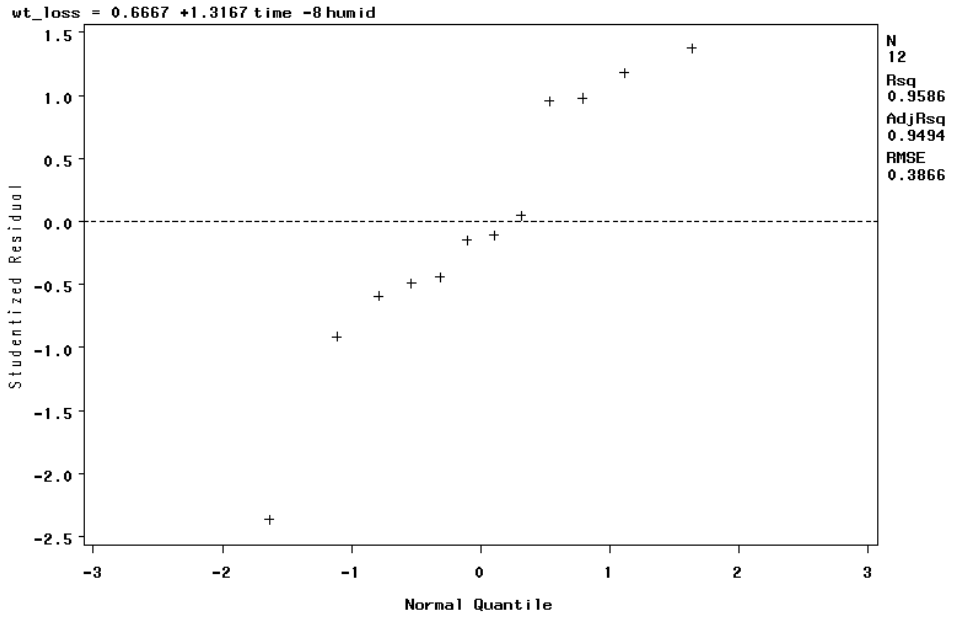
Since $(\beta_1 + \beta_3)$ is the slope for the regression line for drug B compared to the slope of β_1 of

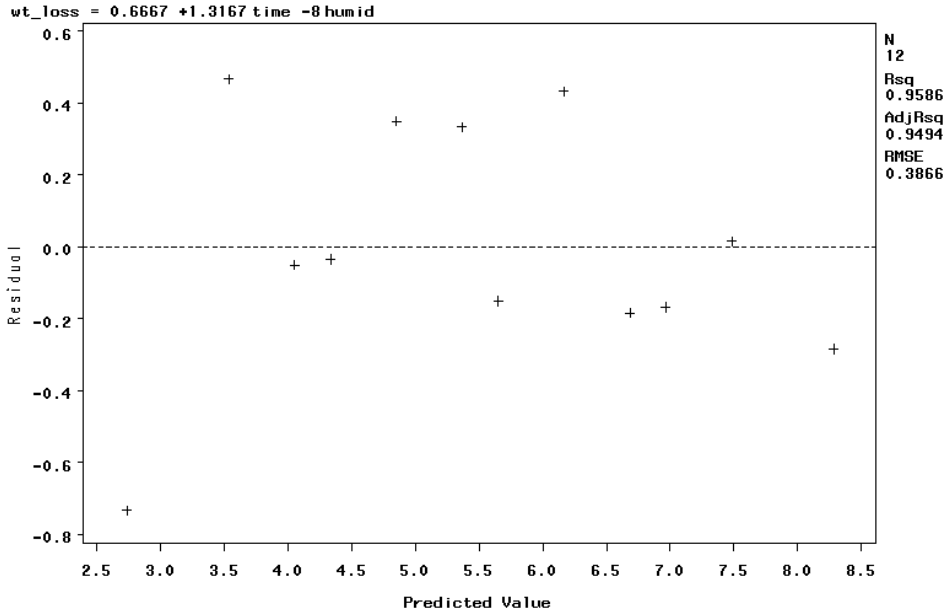
the line for drug A as shown earlier. The fact that $\hat{\beta}_3$ is positive implies that the slope for the line for drug B is larger than that of the line for drug A. Also see the graphic on the left side of Figure 12.4 to visualize this. This implies that *the mean anxiety level increases at a faster rate as the drug dose is increased, for drug product B compared to drug product A.*

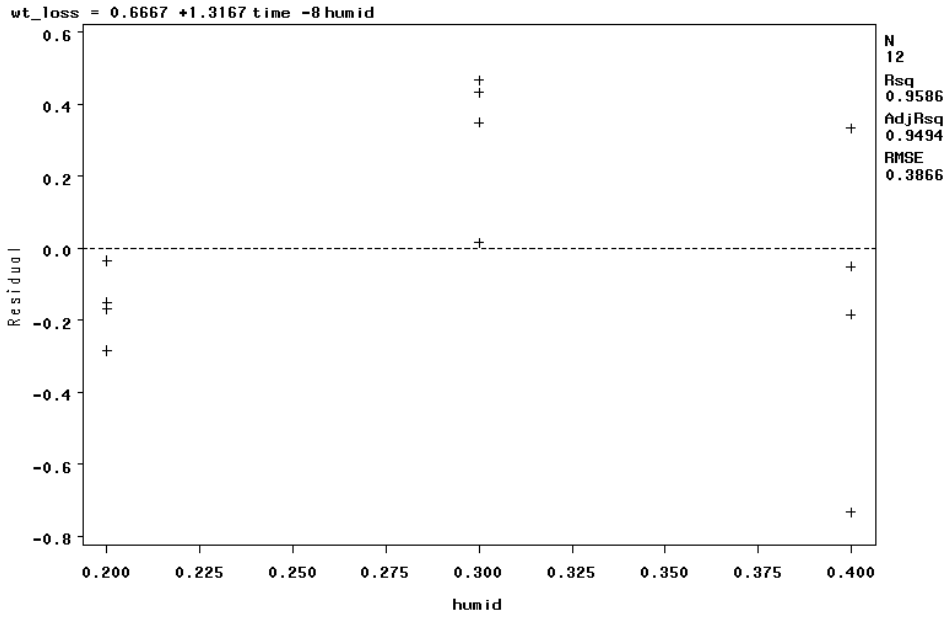
It is important to note that we do not need to use the procedure used above to test for interaction, i.e., $H_0 : \beta_3 = 0$. We could perform this test directly simply using the SAS output for the fit of the model

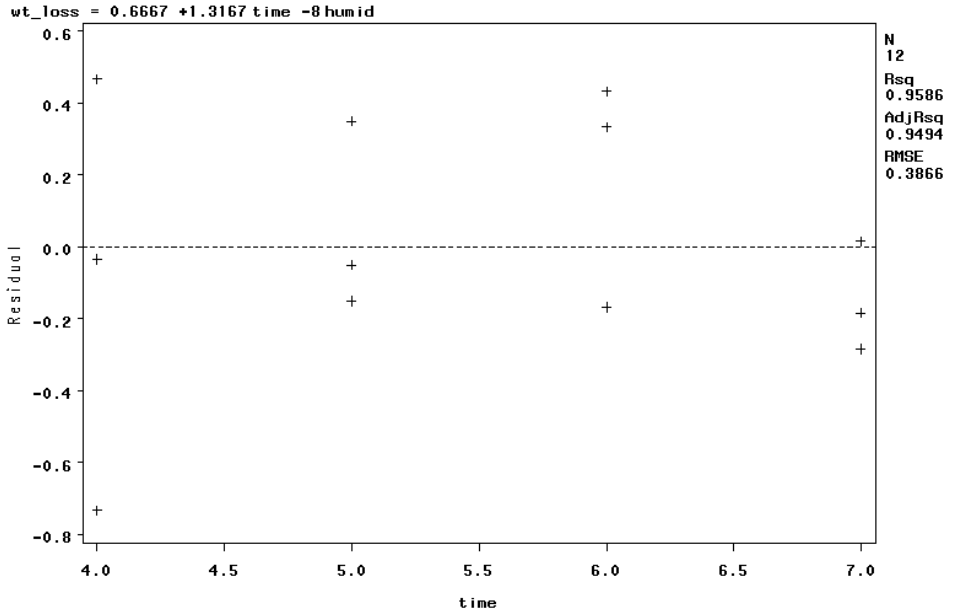
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

The t-statistic (and the p-value) for $x1x2$ term corresponds to the test of the hypothesis $H_0 : \beta_3 = 0$ vs. $H_a : \beta_3 \neq 0$. From the SAS output, the value of this t-statistic is 4.691 and the p-value=.0001. Thus we again reject $H_0 : \beta_3 = 0$ at $\alpha = .05$

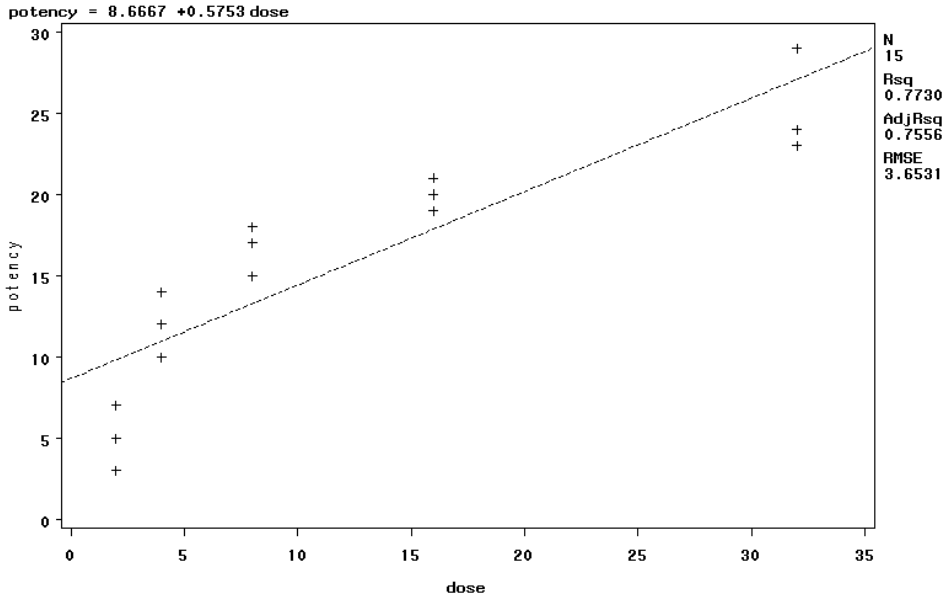


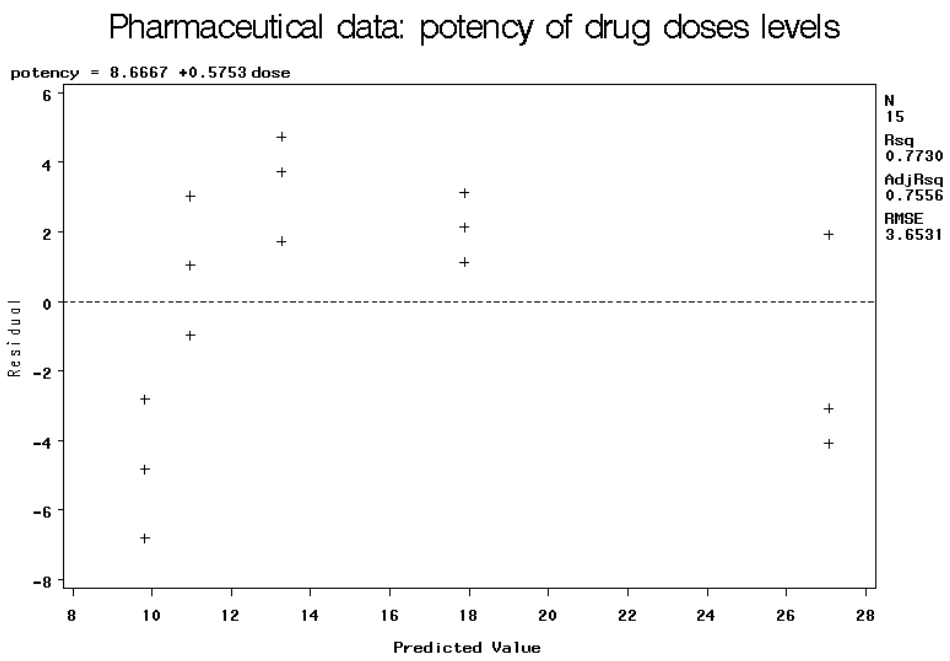

wt_loss = 0.6667 +1.3167 time -8 humid
N
12
Rsq
0.9586
AdjRsq
0.9494
RMSE
0.3866
Residual
Predicted Value

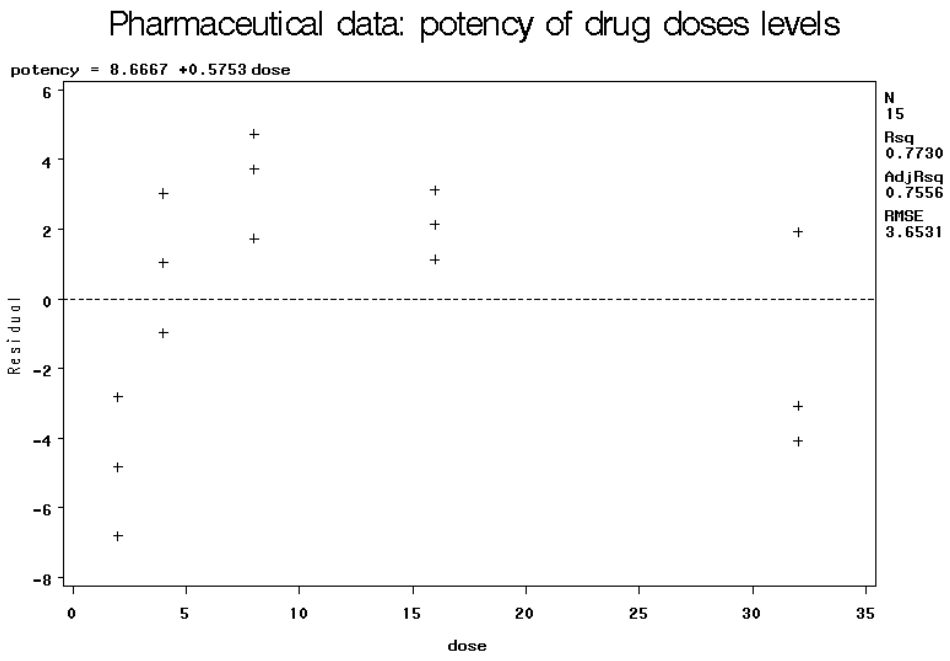

wt_loss = 0.6667 +1.3167 time -8 humid
N
12
Rsq
0.9586
AdjRsq
0.9494
RMSE
0.3866
Residual
humid

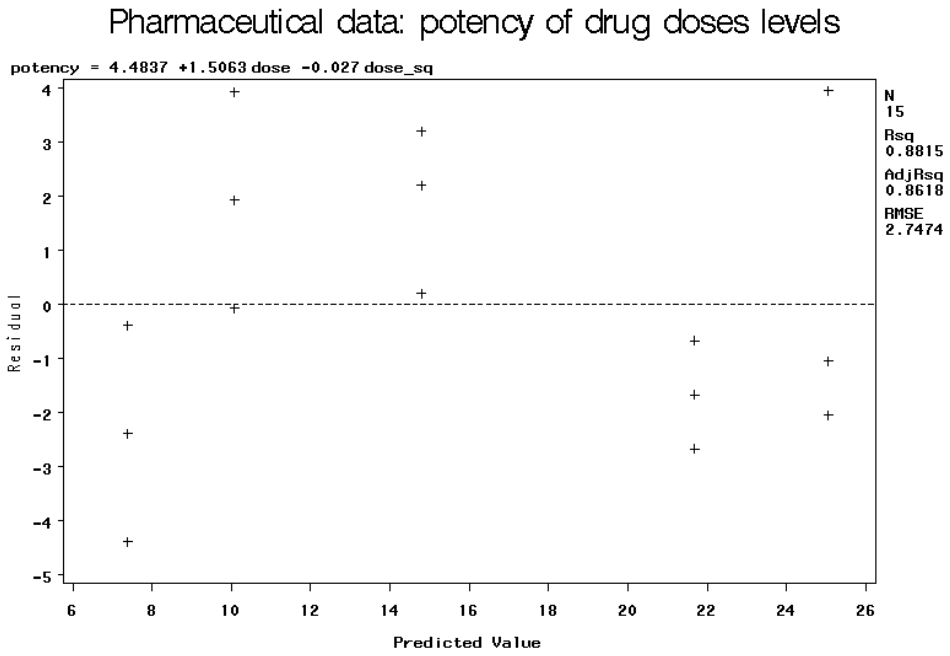


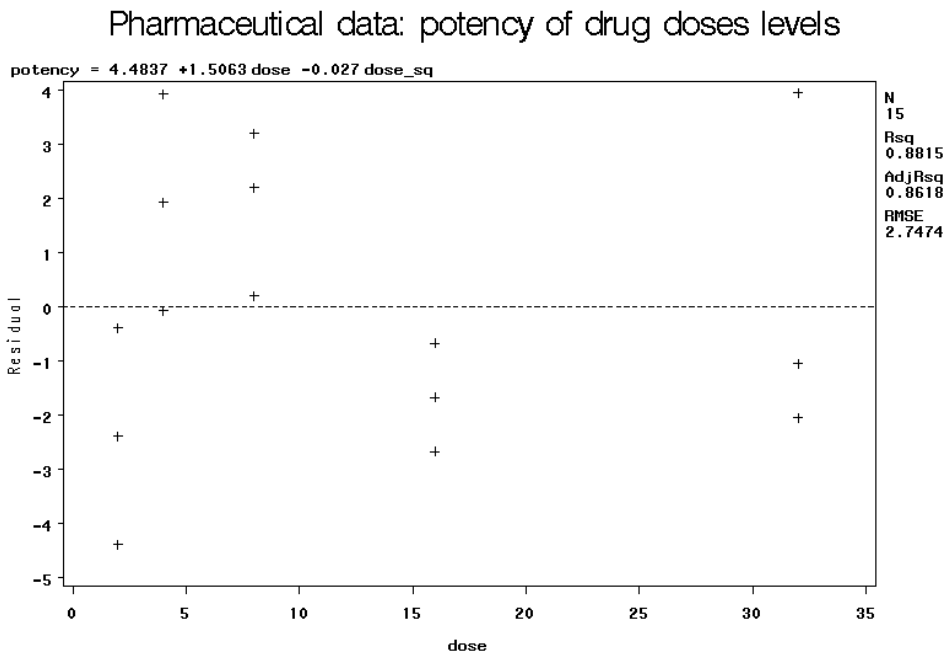
Pharmaceutical data: potency of drug doses levels

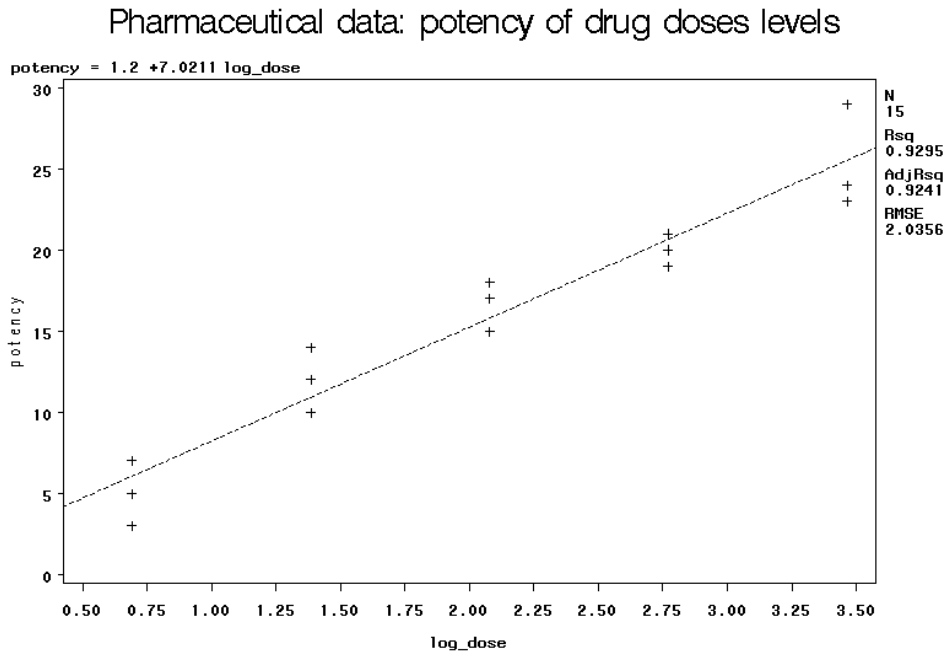


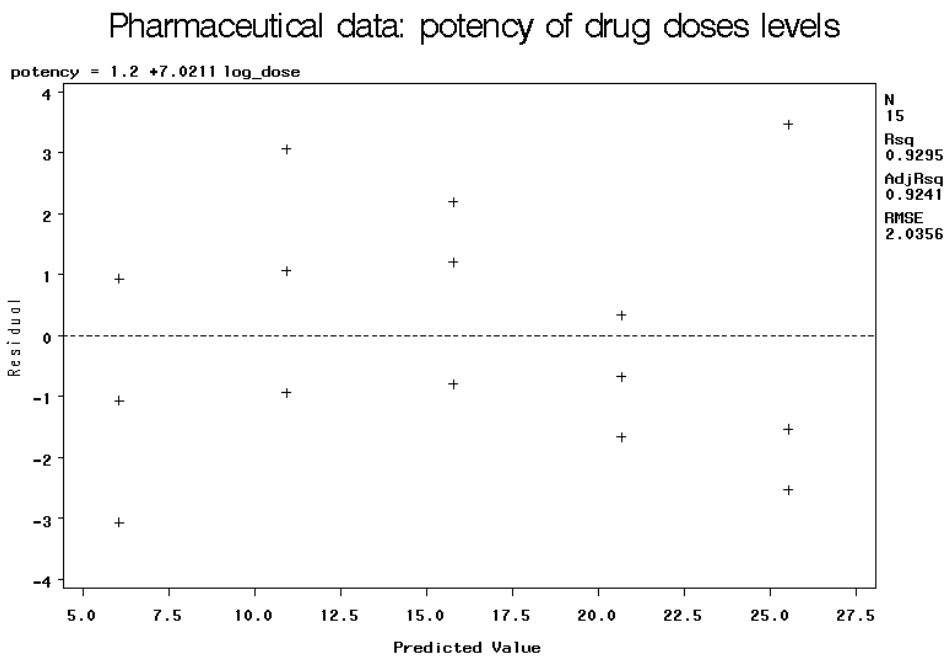

Pharmaceutical data: potency of drug doses levels
potency = 8.6667 +0.5753 dose
N
15
Rsq
0.7730
AdjRsq
0.7556
RMSE
3.6531
Residual
Predicted Value

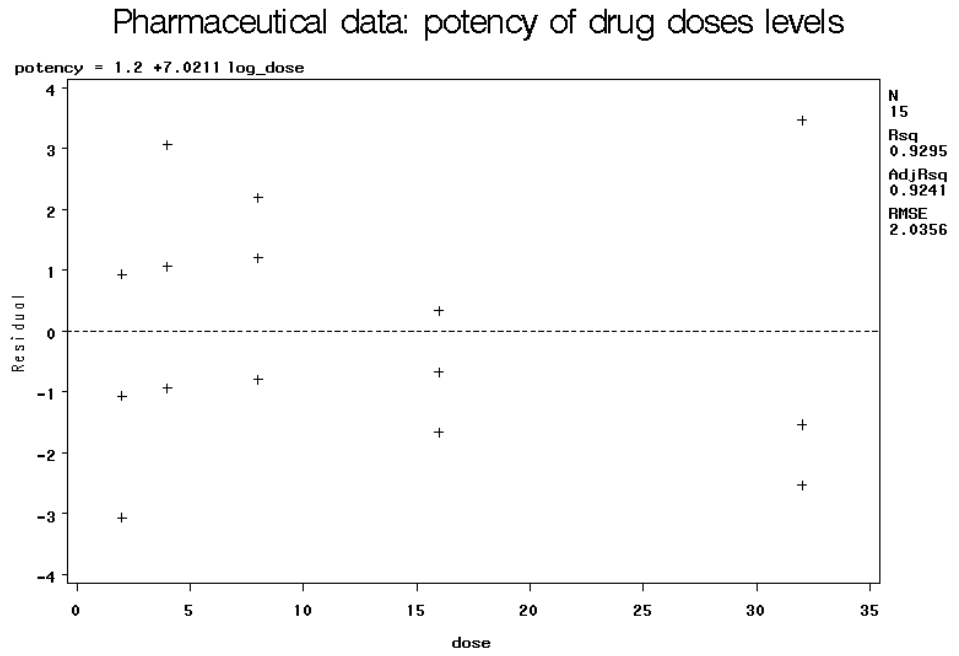

Pharmaceutical data: potency of drug doses levels
potency = 8.6667 +0.5753 dose
N
15
Rsq
0.7730
AdjRsq
0.7556
RMSE
3.6531
Residual
dose


Pharmaceutical data: potency of drug doses levels
potency = 4.4837 +1.5063 dose -0.027 dose_sq
N
15
Rsq
0.8815
AdjRsq
0.8618
RMSE
2.7474
Residual
Predicted Value


Pharmaceutical data: potency of drug doses levels
potency = 4.4837 +1.5063 dose -0.027 dose_sq
N
15
Rsq
0.8815
AdjRsq
0.8618
RMSE
2.7474
Residual
dose


Pharmaceutical data: potency of drug doses levels
potency = 1.2 +7.0211 log_dose
N
15
Rsq
0.9295
AdjRsq
0.9241
RMSE
2.0356
potency
log_dose


Pharmaceutical data: potency of drug doses levels
potency = 1.2 +7.0211 log_dose
N
15
Rsq
0.9295
AdjRsq
0.9241
RMSE
2.0356
Residual
Predicted Value



Multiple Regression in Practical Applications

In Chapters 11–12 we defined and described the computations and statistical inference procedures associated with t linear regression analysis. In this chapter suggestions are given about how to proceed to formulate regression models and progress through an analysis.

Certainly the multiple linear regression tool is used to help accomplish one or more objectives:.

- Often a major objective is to `explain as much of the variation in` Y `as possible`.

Part of the variance in y values is due to the fact that the means of the Y populations are not all the same. By modelling the relationship among the Y population means, we "explain" that part of the overall variance in observed y's which is caused by differences in population means. The remainder of the variance will be "unexplained" or due to random error.

- Certainly, the better the model the more of the variance that will be explained, and that translates into better precision in prediction. So, a second objective – motivating the effort to explain as much variation, is predicting y's at unobserved levels of x's and estimating the Y population mean there.

Naturally, we must select the set of independent variables, the levels of each, and the specific functions of these at which to take observations. Then the observed data are used to refine the model, possibly adding or deleting functions of the x's, to arrive at the most satisfactory

model.

- Often a third objective in multiple regression analysis is simply determining which independent variables are most strongly related to Y, i.e., which ones in combination do the best job of predicting.

Your textbook gives a good description of the steps once usually takes when conducting an analysis. We will follow the outline of the text.

Selecting the Variables(Step 1)

Normally, designating one or more Y variables is natural. These are key variables of interest and learning more about their behavior is the main purpose of the study.

Identifying potentially useful independent variables is not always so easy. Usually, people who are experts in the substantive area of the study will be called on to provide a list of possible x variables. Then data (y) will be collected at selected levels of these x variables.

Usually many more x variables are identified than are actually needed to form a good model. Some subsets of the x's are often so closely related to one another that each provides about the same information about y, hence only one variable from the subset is needed in the model. Identifying such redundancies is the major challenge in variable selection.

One selection procedure which may be used is to do `all possible regressions`. This means fit every possible submodel of the full model involving all independent variables. Then a "good" submodel is selected based on some criterion related to the model fits. Some criteria often used are:

1. Small s_ϵ^2
2. Large R^2

Unfortunately, the full model always gives the smallest s_ϵ^2 and the largest R^2. What is a "good" submodel will be one that is a compromise between the desires to optimize s_ϵ^2 and R^2, and use as few x variables as possible. Usually more than one "good" submodel will emerge, and a decision to select one must somehow be made.

One technique that is sometimes used to help in this regard is `data splitting`. This means dividing the data into two parts. The first part is used to compute all possible regressions (or

some other method). The second part of the data are used to see how well various submodels predict "future" observations, i.e., the y's in this second subset. This technique can help sort out the better of the "good" submodels. Most often the data are split by dividing them in half.

If there are too few observations to split, a statistic named `PRESS` may be used. For each of the "good" submodels do the following. For each of the n data elements in turn, delete the element, fit the model using the $n-1$ remaining data elements, and use the fit to predict the y of the deleted observation. Call the prediction $\hat{y}_i^*$ and y_i the observation deleted. Then

$$\text{PRESS} = \sum_{\text{i}=1}^{\text{n}} (\text{y}_\text{i} - \hat{\text{y}}_\text{i}^*)^2 .$$

Submodels with `small` PRESS are preferable. PRESS may be used in conjunction with all possible regressions to decide between two or more "good" submodels, or PRESS can be used alone to help select a useful subset of the x's.

`It must be emphasized that a ''best'' subset of independent variables is not defined so doesn't exist`. We are looking for "good" subsets, and will then select one of these as our model.

A competitor technique to all possible regressions and PRESS is the use of the C_p statistic. It is defined as

$$C_p = \frac{\text{SSE}_\text{p}}{s_\epsilon^2} - (n - 2p)$$

Where SSE_p is the error sum of squares due to fitting the submodel containing p parameters (including β_0), and s_ϵ^2 is the error SS due to fitting all available x variables. "Good" submodels have C_p which is nearly p, i.e., $C_p \approx p$. This technique is especially good at helping to avoid selecting `overspecified` model having redundant x variables.

There are some numerical algorithms available which automate the subset selection process. Some people call them "best subset selection algorithms", but they cannot possibly select the "best" because it isn't defined. Nevertheless these algorithms are used and they often select reasonably good subsets. Some of these are:

1. Backward Elimination

This algorithm begins with the full model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

and attempts to delete a variable x_j to obtain a model with one less independent variable. The variable selected for deletion is the one that causes least increase in Error SS of the full model. It can be shown that this is equivalent to finding the smallest `F-to-delete` statistic

$$F_j = \frac{\mathrm{SS_{DROP_j}}}{\mathrm{MSE}} \qquad j = 1, 2, \ldots$$

where MSE is the mean square error from the full model fit, and the numerator is the difference in Error SS of the two models. If this F-statistic is less than a preselected cut-off value the corresponding x variable is deleted from the model.

The cut-off value is usually an F-value $F_\alpha, 1, \mathrm{df}$, where α is a preset value (usually called the significance level for deletion) and df= d.f. for MSE. The process continued until a variable fails to be deleted at which time the x variables ramaining in the model comprise the selected subset.

2. Forward Selection

The algorithm begins with the model $y = \beta_0 + \epsilon$, and extends model by attempting to add one term at a time. The variable selected at each stage is the one that causes largest decrease in Error SS of the full model. This is equivalent to finding the largest F-to-enter statistic

$$F_j = \frac{\mathrm{SS_{ADD_j}}}{\mathrm{MSE}} \qquad j = 1, 2, \ldots$$

where MSE is the mean square error from the full model fit, and the numerator is the difference in Error SS of the two models. If this F-statistic is greater than a preselected cut-off value the corresponding x variable is added to the model.

The cut-off value is usually an F-value $F_\alpha, 1, \mathrm{df}$, where α is a preset value (usually called the significance level for entry) and df= d.f. for MSE. When a variable fails to enter, the process stops and the submodel which exist at that time is declared the model selected by the procedure.

3. Stepwise Regression

This algorithm combines the steps of forward selection and backward elimination. It begins with the model $y = \beta_0 + \epsilon$ and does a forward selection step. If a variable is added then the current model is (if x_j is added)

$$y = \beta_0 + \beta_1 x_j + \epsilon$$

Now a backward elimination (modified) step is made. If an x variable is deleted then the current model has one fewer terms. Otherwise it is unchanged and a forward selection step is made.

Thus alternating forward and backward steps are made until no variable is added to the current subset in a forward step and no variabale is deleted in the following backward step. Then the algorithm stops and the current subset at that time comprise the selected model.

Any or all of the methods mentioned can be used to help decide which set(s) of independent variables will be used. It isn't necessary to decide absolutely on only one. We can go forward with more than one, and make a final decision at a later stage in the overall analysis.

Model Formation (Step 2)

Step 1 yields a subset $x_1, x_2, \ldots, x_k$ of the original full set of independent variables (sometimes we will proceed to Step 2 with the full set).

The second step is to decide exactly which terms will be in the model. For example, if we have x_1, x_2, x_3, do we want the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

or $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \epsilon$$

or $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_3 + \epsilon$$

or $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \epsilon$$

$\vdots$

There are many other possibilities.

A reasonable way to proceed (most people do this) is to fit the model which involves only first order terms, i.e.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

to get residuals

$$y_i - \hat{y}_i$$

Plotting residuals versus each x may then suggest which higher-degree terms may be appropriate. Then trying various higher order term model fits may lead to desirable extensions of the first order model.

Plotting will not, however, suggest interaction between more than two variables. If such are suspected, a model having many higher order terms, some of them higher order interactions,

can be subjected to the `stepwise regression` algorithm to see which terms are retained. Any of the other algorithms (forward or backward) might be used instead.

Also, if repeat y observations have been made at one or more combination of independent variable levels, the test for lack of fit can be used to identify model which may be improved by adding higher order terms. *This is not an absolute because lack of fit may be due to the need for nonlinear functions of the parameters.* We won't consider nonlinear terms and models in Stat. 401.

Remember, there is usually no "best" model because there is usually no exact functional relationship among Y population means. Thus different researchers may well find different very good models for the same data.

Checking model assumptions (Step 3) by analyzing residuals

The assumptions we have made are all about the ϵ_i in the model. They are:

1. $E(\epsilon_i) = 0$ all i
2. $V(\epsilon_i) = \sigma^2_\epsilon$, all i
3. ϵ_i is normally distributed, all i
4. The ϵ_i are independent

We already discussed some graphical procedures for residuals from the simple linear regression model in Chapter 11.

Since the residuals $r_i = y_i - \hat{y}_i$ are estimates of the ϵ_i, they are used in our efforts to detect violations.

1. $E(\epsilon_i) = 0$ (Model Adequacy)

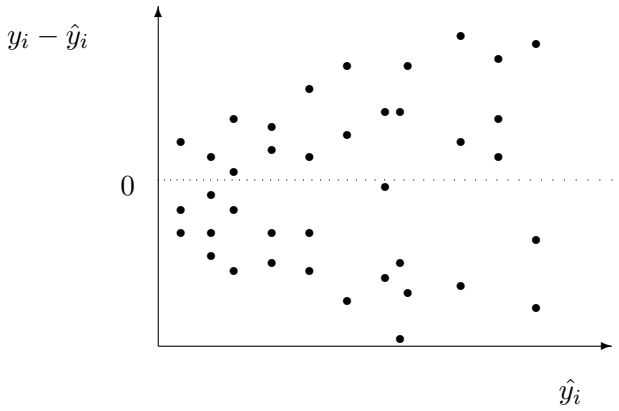
If a linear regression model fits well to the data, this assumption won't be seriously violated, in which case we would expect the residuals to be evenly spaced around the zero line when $r_i = y_i - \hat{y}_i$ are plotted against each of the independent variables x_i. A noticeable curvature or nonlinearity is present in this plot may indicate that a more complicated relationship such as a polynomial. In this case a tranformation of the y variable or the partcular x_i variables

may be attempted.

2. $V(\epsilon_i) = \sigma_\epsilon^2$ **(Constant Error Variance)**

Plotting residuals $r_i = y_i - \hat{y}_i$ versus predicted values $\hat{y}_i$ will show significant heterogeneity if it exists, as shown below. Nonconstant variance will make the width of confidence and prediction intervals too wide or too narrow depending on the variance at the considered x value.

Transformations to stabilize variance are possible, and a technique called weighted least squares can be used to overcome some of the adverse effects of heterogeneity of variance. We won't pursue these topics now. A plot suggesting heterogeneity is:



Pattern Showing Heterogeneity of Variance

3. ϵ_i **Normal**

A Normal probability plot of the residuals or better the `studentized residuals` $(y_i - \hat{y}_i)/s_\epsilon$ versus percentiles from the standard normal distribution. If this plot shows a substantial deviation from linearity the normality assumption about ϵ's is seriously in doubt. Particularly one you should look for one of the pattern that indicate that the error disribution is deviates from a Normal distribution in one of the ways we discussed in class.

Also Histogram plots and Box plots of residuals can show skewness in distribution which indicates nonnormality, (lack of symmetry to be specific).

Serious violation of this assumption does not impact point estimate, but distributional assumptions for t and F tests are not valid and these test may be seriously invalidated. An available remedy is again try a transformation of the y variable.

4. Independent ϵ_i (Uncorrelated)

This one is hard to check. We will talk about a check only when the observations have a time-oriented sequence. For observations taken over time we talk about serial correlation which measures the tendency of successive adjacent ϵ's to be similar (positive serial correlation) or dissimilar (negative serial correlation). Independence gives zero serial correlation.

A statistic which estimates serial correlation is the Durbin-Watson statistic. Let $\hat{\epsilon}_t$ be the residual for the observation at time t. Then

$$d = \sum_{t=1}^{n-1} (\hat{\epsilon}_{t+1} - \hat{\epsilon}_t)^2 \ / \ \sum \hat{\epsilon}_t^2$$

when there is no serial correlation $E(d) = 2.0$. When the observed d value is outside the interval (1.5, 2.5) suggests nonzero serial correlation.

A good idea is to try to avoid having dependent ϵ_i when collecting data. Introduce randomness as much as possible in the data collection process and try to avoid taking observations where one depends on the value obtained for another.