Model Building

Hong Kong Baptist University

Fall 2021

Model Building

Model building means writing a model that will provide a good fit to a set of data and that will give good estimates of the mean response and good predictions of the response for given values of the independent variables.

Terms have to be added to the model to account for interrelationships among the independent variables and for curvature in the response function Y *.*

Failure to include needed terms causes (a) inflated values of SSE, (b) insignificance in statistical tests, and, often, (c) erroneous practical conclusions.

Continuous (quantitative) independent variables are treated differently from categorical (qualitative) independent variables.

Polynomial Regression

First, consider the polynomial regression model for one continuous independent variable x:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon. \tag{1}$$

To decide the order k in the model building process, we should first construct a scatterplot. We know that a k^{th}-order polynomial, when graphed, will exhibit $k-1$ peaks, troughs or reversals in direction.

In real applications, most responses are curvilinear and so we should try a second-order model in order to capture the curvature.

Third- or higher-order models would be used only when you expect more than one reversal in the direction of the curve. These situations are rare, except where the response is a function of time.

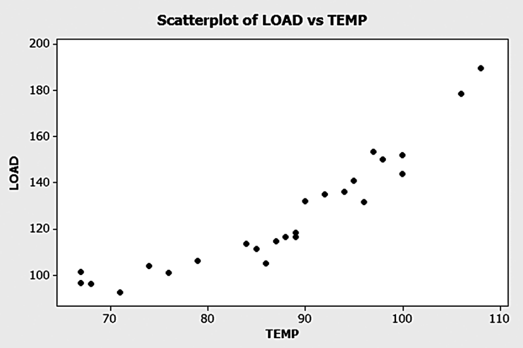
Example 5.2

Table 5.1 Power load data

Temperature °F	Peak Load megawatts	Temperature °F	Peak Load megawatts	Temperature °F	Peak Load megawatts
94	136.0	106	178.2	76	100.9
96	131.7	67	101.6	68	96.3
95	140.7	71	92.5	92	135.1
108	189.3	100	151.9	100	143.6
67	96.5	79	106.2	85	111.4
88	116.4	97	153.2	89	116.5
89	118.5	98	150.1	74	103.9
84	113.4	87	114.7	86	105.1
90	132.0				

5- 13

Figure 5.5 MINITAB scatterplot for power load data



5- 14

Figure 5.6 MINITAB output for third-order model of power load

```
The regression equation is
LOAD = 331 - 6.4 TEMP + 0.038 TEMP2 +0.000084 TEMP3

Predictor        Coef     SE Coef        T        P
Constant        331.3       477.1     0.69    0.495
TEMP            -6.39       16.79    -0.38    0.707
TEMP2          0.0378      0.1945     0.19    0.848
TEMP3       0.0000843   0.0007426     0.11    0.911

S = 5.501       R-Sq = 95.9%     R-Sq(adj) = 95.4%

Analysis of Variance

Source          DF        SS        MS        F       P
Regression       3   15012.2    5004.1   165.36   0.000
Residual Error  21     635.5      30.3
Total           24   15647.7
```

Figure 5.7 MINITAB output for second-order model of power load

The regression equation is
LOAD = 385 - 8.29 TEMP + 0.0598 TEMP2

Predictor	Coef	SE Coef	T	P
Constant	385.05	55.17	6.98	0.000
TEMP	-8.293	1.299	-6.38	0.000
TEMP2	0.059823	0.007549	7.93	0.000

S = 5.376   R-Sq = 95.9%   R-Sq(adj) = 95.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	15011.8	7505.9	259.69	0.000
Residual Error	22	635.9	28.9		
Total	24	15647.7			

- Example 5.2 gives us an example where the second-order polynomial model (quadratic regression) is significantly better than the first-order model (simple linear regression) and is not significantly worse than the third-order model (cubic model).
- By saying ‘significantly better’, we mean the parameter for the quadratic term in the quadratic model is significant (but it does not necessarily mean that the parameter for the quadratic term in the cubic model or in even higher order models is still significant).
- By saying ‘not significantly worse’, we mean the parameter for the cubic term in the cubic model is not significant.

- R^2 increases when more terms are included in the model but the increased amount can be small, while the adjusted R^2 does not necessarily increase together with the number of parameters.
- The adjusted R^2 of the quadratic regression model is higher than that of the cubic model, which has a higher R^2 than the quadratic regression.
- Such a comparison of the (adjusted) R^2 values among the models for helping us decide whether a term should be included or not will be discussed in more details in Chapter 6.

- For the cubic regression, we also encountered in this numerical example a situation that the p-value in the ANOVA is small, suggesting that not all coefficients are zero, whilst the p-values in the t-tests for testing individual coefficients are all large, suggesting that each coefficient is not significant, when tested individually.
- Such a paradoxical situation is *not* due to the inflated overall type I error rate in multiple testing (the t-tests do not suggest rejection and so we will not commit type I error!), but due to the problem of **multicollinearity**, which will be discussed in more details later in Chapter 7, but the next page gives a brief explanation of the phenomenon observed in the ANOVA and t-tests.

- Technically speaking, multicollinearity happens when the column vectors in $\boldsymbol{X}$ are not linearly independent.
- E.g. when $x_2 = 2x_1$ in $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, then the true model is in fact $Y = \beta_0 + \beta_1^* x_1 + \varepsilon$. However, in the model with both x_1 and x_2, we actually split the single parameter β_1^* into two parameters β_1 and β_2 such that $\beta_1 + 2\beta_2 = \beta_1^*$, and then we estimate these two parameters; under such a situation, either β_1 or β_2 alone will have an infinite standard deviation (because any one of them alone can be any value) and hence is not significant in the individual t-test, but if β_1^* is significant, then β_1 and β_2 cannot be both insignificant in the F-test.

First-order model with k continuous independent variables

The above polynomial regression model is for one continuous independent variable. Now, suppose we have k continuous independent variables.

We may simply form the first-order model, and then the response surface is just a hyperplane in a $(k+1)$-dimensional space, i.e. there is no curvature and the contour lines are parallel. That is, the independent variables affect the response independently of each other and so the independent variables do not interact.

First-Order Model in k Quantitative Independent Variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters that must be estimated.

Interpretation of model parameters

β_0: y-intercept of $(k+1)$-dimensional surface; the value of $E(y)$ when $x_1 = x_2 = \cdots = x_k = 0$

β_1: Change in $E(y)$ for a 1-unit increase in x_1, when $x_2, x_3, \ldots, x_k$ are held fixed

β_2: Change in $E(y)$ for a 1-unit increase in x_2, when $x_1, x_3, \ldots, x_k$ are held fixed

$\vdots$

β_k: Change in $E(y)$ for a 1-unit increase in x_k, when $x_1, x_2, \ldots, x_{k-1}$ are held fixed

Figure 5.8 Response surface for first-order model with two quantitative independent variables

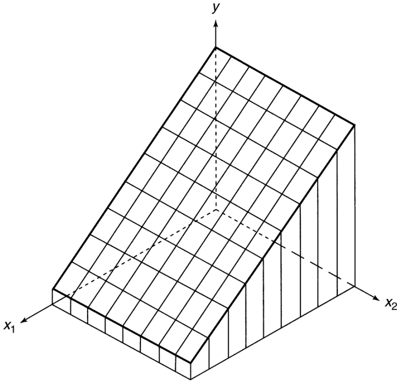
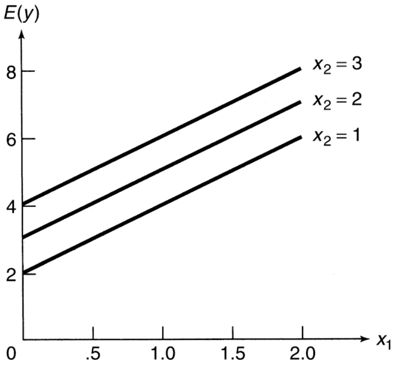


Figure 5.9 Contour lines of $E(y)$ for x_2 = 1,2,3 (first-order model)



5- 20

Interaction (second-order) model with k continuous independent variables

If we include the second-order interaction terms (i.e. the products $x_i x_j$), then e.g. for $k = 2$, the response surface is a twisted plane, which can be obtained by twisting (but not bending or folding) a sheet of paper.

Consequently the contour lines are nonparallel, meaning that the effect of a one-unit change in one independent variable, while keeping the other independent variables fixed, will depend on the values of the other independent variables, i.e. the slope (and the intercept) depends on the values of the other independent variables.

Interaction (Second-Order) Model with Two Independent Variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Interpretation of Model Parameters

β_0: y-intercept; the value of $E(y)$ when $x_1 = x_2 = 0$

β_1 and β_2: Changing β_1 and β_2 causes the surface to shift along the x_1 and x_2 axes

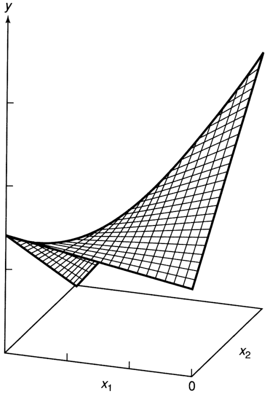
β_3: Controls the rate of twist in the ruled surface (see Figure 5.10)

When one independent variable is held fixed, the model produces straight lines with the following slopes:

$\beta_1 + \beta_3 x_2$: Change in $E(y)$ for a 1-unit increase in x_1, when x_2 is held fixed

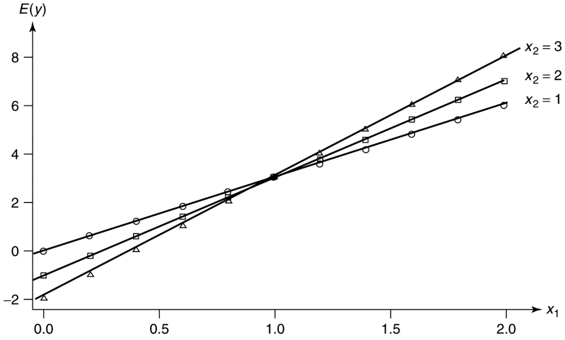
$\beta_2 + \beta_3 x_1$: Change in $E(y)$ for a 1-unit increase in x_2, when x_1 is held fixed

Figure 5.10 Response surface for an interaction model (second-order)



5- 22

Figure 5.11 Contour lines of $E(y)$ for x_2 = 1,2,3 (first-order model plus interaction)



5- 23

Complete second-order model with k continuous independent variables

If we include all second-order interaction terms and second-order terms of individual variables (i.e. the squares x_i^2), we have the complete second-order model.

For $k = 2$, the three possible response surface are a paraboloid opening upward, a paraboloid opening downward and a saddle-shaped surface.

The response surfaces for higher-order models would have very complicated geometrical structures, and in real applications, we seldom consider third- or higher-orders unless there are good scientific reasons to expect more than one reversals in direction.

Complete Second-Order Model with Two Independent Variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

Interpretation of Model Parameters

β_0: y-intercept; the value of $E(y)$ when $x_1 = x_2 = 0$

β_1 and β_2: Changing β_1 and β_2 causes the surface to shift along the x_1 and x_2 axes

β_3: The value of β_3 controls the rotation of the surface

β_4 and β_5: Signs and values of these parameters control the type of surface and the rates of curvature

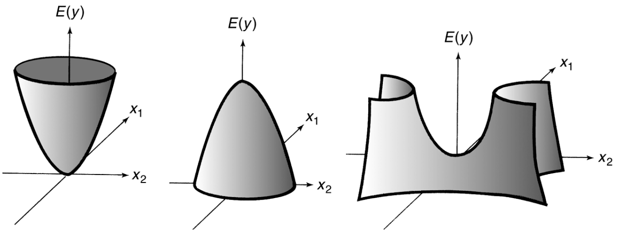
Three types of surfaces may be produced by a second-order model.*

A paraboloid that opens upward (Figure 5.12a)

A paraboloid that opens downward (Figure 5.12b)

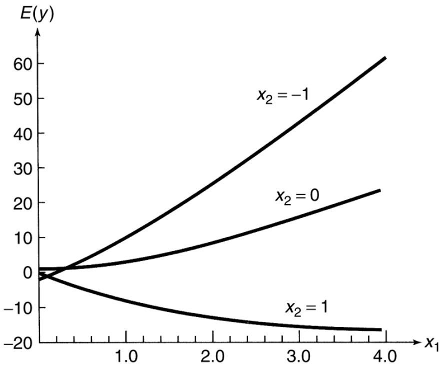
A saddle-shaped surface (Figure 5.12c)

Figure 5.12 Graphs of three second-order surfaces



(a) (b) (c)

Figure 5.13 Contours of $E(y)$ for x_2 = -1,0,1 (complete second-order model)



Coding of a continuous independent variable

For a continuous independent variable, say x_i, we also have to consider whether we should standardise (or normalise) it by

$$u_{ij} = \frac{x_{ij} - \overline{x}_i}{s_i},$$

where $\overline{x}_i$ and s_i are the sample mean and sample standard deviation of the observed values $\{x_{i1}, x_{i2}, \ldots, x_{in}\}$ of the independent variable x_i. The resultant variable is then denoted by u_i. Such a standardisation (normalisation) is a kind of *coding* the independent variables. The advantages of standardisation are

(i) the new standardised origin = the centre of the standardised values,

(ii) the range of u_{ij} is approximately the same (mostly between -3 and $+3$) for each fixed i,

(iii) the correlation between x_i and x_i^2, after standardisation, i.e. between u_i and u_i^2, will be reduced (we will explain why below).

Why these properties are advantages? Because of the two potential problems:

- Rounding error: considerable rounding error may occur in the computation of the inverse of the information matrix $\boldsymbol{X}'\boldsymbol{X}$, if the numbers in the matrix vary greatly in absolute value. Thus, points (i) and (ii) above help us cope with the problem of rounding error.
- Multicollinearity: When polynomial regression models are used, the problem of multicollinearity is unavoidable, especially when higher-order terms are included. The likelihood of rounding errors in the regression coefficients is increased in the presence of these highly correlated independent variables. Point (iii) above reduces the trouble caused by multicollinearity.

However, the interpretation of the least squares estimates for standardised variables is indirect (and may be difficult to be understood by laymen).

A one-unit change in u is equal to a one-sample standard deviation change in x, and so for a simple linear regression, a one-unit change in x is accompanied by a $(\frac{\hat{\beta}}{s_x})$-unit change in the mean of Y.

However, the sample deviation is nothing but just a numerical value from the sample, not any universal constant having physical meaning. Thus, it is likely not understandable to laymen if one says “increasing x by one sample standard deviation”.

On the other hand, the interpretation of the intercept term after standardisation may be more meaningful than that in the original model.

In the original model, the intercept is the estimated mean response when all independent variables are zero, but in some applications zero-valued independent variables (e.g. height, weight) are meaningless and so is the intercept.

In the standardised model, the intercept is the estimated mean response when all standardised independent variables are zero, i.e. when all independent variables (without standardisation) are set at their mean values in the sample; hence there is no extrapolation and the intercept is always meaningful. [This interpretation of the intercept is a consequence of point (i) above.]

For a polynomial regression model, there are two possible ways to standardise higher-order terms x^k, namely, we either take the k^{th} power of the standardised value u or standardise directly the values x^k.

The former will allow us to have again a polynomial regression model but the computational advantages mentioned in points (i) and (ii) above are no longer true [nevertheless, point (iii) is actually referring to such a procedure and so it remains true].

The latter will create a regression model that is no longer a polynomial, leading to entirely different interpretations, and so is not recommended.

Now, why point (iii) is true? Let's start with an intuitive explanation first.

Clearly, if x is only on the positive half line, when x increases, then x^2 increases, and so they are positively correlated, and if x is only on the negative half line, the correlation is negative.

To visualize it, consider the curve $f(x) = x^2$. If you focus only on the positive part (or negative part, respectively), you will see a positive correlation (or negative correlation, respectively), and the correlation is getting more positive (or more negative, respectively) when you move the interval to the right-hand side (or to the left-hand side, respectively); that is, the further away from zero the interval of x is, the stronger the correlation between x and x^2 over that interval it will be.

By centering, i.e. $x_* = x - \overline{x}$, you move the interval so that the mid-point of the interval is zero, and consequently the correlation will be weakened, because then the correlation is negative on the left half of the interval and positive on the right half.

Also, after centering, we are considering the flattest part of the curve (as just mentioned above, the correlation is small when the interval is close to zero).

Thus, when we consider the correlation between x_* and x_*^2, this correlation should be lower than the original correlation between x and x^2.

More mathematically, after some algebra, it can be shown that

$$\text{sample cov}(x, x^2) = m_3 + 2m_2\overline{x},$$

where $m_k = \sum_{i=1}^{n}(x_i - \overline{x})^k/(n-1)$ is the k^{th} centered sample moment of x.

Because m_k is the centered moment, shifting x_i by the same amount will not change m_k, i.e. m_k for x is also m_k for x_*.

Thus, if x_i are positive (or negative, respectively), then $2m_2\overline{x}$ is positive (or negative, respectively), and after centering,

$$\text{sample cov}(x_*, x_*^2) = m_3$$

is closer to zero than sample cov(x, x^2), and hence the severity of the problem of multicollinearity will be reduced.

This mathematics also suggests an even better coding, namely $x_{**} = x -$ something, so that $\overline{x}_{**} = -m_3/(2m_2)$, leading to a zero sample covariance between x_{**} and x_{**}^2. This results in the so-called *orthogonal polynomial*, a topic that is beyond the scope of this course.

However, it should be kept in mind that standardisation is an option but not a must; whether you standardise or you don't standardise is a matter of personal judgement.

In SAS, the procedure `proc standard` can be used to standardise the variable(s) specified after `var`] and the standardised values are still under the same variable but in a new data set whose name is given after `out=`, `zmos` say.

To create new variables such as the square of the standardised variable, we have to use `data` again to make another new data set. The command `set zmos` will move all variables in the data set `zmos` to the new data set, and because we are under `data`, we are allowed to transform existing variables to create new variables.

The option `corr` after `proc reg` will report the correlations between all variables in the `model` statement(s).

The procedure `proc corr` is different from the option `corr` of `proc reg`; it is a procedure for calculating correlations between variables, and the p-values for testing zero correlation are reported for each correlation.

Coding Procedure for Observational Data

Let

$$x = \text{Uncoded quantitative independent variable}$$

$$u = \text{Coded quantitative independent variable}$$

Then if x takes values $x_1, x_2, \ldots, x_n$ for the n data points in the regression analysis, let

$$u_i = \frac{x_i - \bar{x}}{s_x}$$

where s_x is the standard deviation of the x-values, that is,

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

Table 5.3 Data for Example 5.4

Date	Average Temperature, x	Catch Ratio, y
July 24	16.8	.66
25	15.0	.30
26	16.5	.46
27	17.7	.44
28	20.6	.67
29	22.6	.99
30	23.3	.75
31	18.2	.24
Aug. 1	18.6	.51

Source: Petric, D., et al. "Dependence of CO_2-baited suction trap captures on temperature variations," *Journal of the American Mosquito Control Association*, Vol. 11, No. 1, Mar. 1995, p. 8.

Figure 5.16 MINITAB printout for the quadratic model, Example 5.4

Regression Analysis: RATIO versus TEMP, TEMPSQ

```
The regression equation is
RATIO = 1.09 - 0.119 TEMP + 0.00471 TEMPSQ


Predictor       Coef    SE Coef      T      P
Constant       1.091      3.380   0.32  0.758
TEMP         -0.1186     0.3537  -0.34  0.749
TEMPSQ      0.004705   0.009103   0.52  0.624


S = 0.170451   R-Sq = 60.4%   R-Sq(adj) = 47.2%


Analysis of Variance

Source           DF       SS       MS     F      P
Regression        2  0.26563  0.13282  4.57  0.062
Residual Error    6  0.17432  0.02905
Total             8  0.43996
```

Correlations: TEMP, TEMPSQ

```
Pearson correlation of TEMP and TEMPSQ = 0.998
P-Value = 0.000
```

Table 5.4 Coded values of x, Example 5.4

Temperature, x	Coded Values, u
16.8	−.71
15.0	−1.36
16.5	−.82
17.7	−.39
20.6	.64
22.6	1.36
23.3	1.61
18.2	−.21
18.6	−.07

Figure 5.18 MINITAB printout for the quadratic model with coded temperature

Correlations: U, USQ

```
Pearson correlation of U and USQ = 0.441
P-Value = 0.235
```

Regression Analysis: RATIO versus U, USQ

```
The regression equation is
RATIO = 0.525 + 0.164 U + 0.0372 USQ
```

Predictor	Coef	SE Coef	T	P
Constant	0.52469	0.08558	6.13	0.001
U	0.16423	0.06713	2.45	0.050
USQ	0.03721	0.07198	0.52	0.624

```
S = 0.170451   R-Sq = 60.4%   R-Sq(adj) = 47.2%
```

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	0.26563	0.13282	4.57	0.062
Residual Error	6	0.17432	0.02905		
Total	8	0.43996			

In Example 5.4 we can see the correlation between `temp` and `temp2` is much higher than that between `u` and `u2`, while the F-test in ANOVA, the MSE, R^2, etc. all remain the same after standardisation.

For individual t-tests, the p-value for the highest order term will remain the same because its coefficient will only be rescaled by dividing by the sample standard deviation after standardisation; asking whether the rescaled parameter is equal to zero is the same as asking whether the original one is equal to zero.

However, for a lower-order term, after centring the coefficient will not only be rescaled but actually will be changed to a linear combination of several parameters in the original model, and so testing whether the coefficient of a lower-order standardised term is zero is different from testing whether the original coefficient is zero. (Recall that the intercept, which is a lower-order term, is interpreted differently in the original model and the standardised model.)

Categorical independent variables

One-way ANOVA

For a regression model with only one independent variable, which is a categorical variable having m levels, there is only one model

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{m-1} x_{m-1} + \varepsilon,$$

where x_i are the dummy variables.

Note that $x_i^k = x_i$ for any positive integer k and $x_i x_j = 0$ whenever $i \neq j$. Thus, we do not have any "higher-order terms" x_i^k or "interaction terms" $x_i x_j$.

This is exactly the one-way ANOVA model, and testing whether $\beta_1 = \cdots = \beta_{m-1} = 0$ is the same as testing whether $\mu_1 = \cdots = \mu_m$, where μ_i is the mean response at level i.

Table 5.6 Annual maintenance costs

	State Installation		
	Kansas	Kentucky	Texas
	$ 198	$ 563	$ 385
	126	314	693
	443	483	266
	570	144	586
	286	585	178
	184	377	773
	105	264	308
	216	185	430
	465	330	644
	203	354	515
Totals	$2,796	$3,599	$4,778

Figure 5.19 SPSS printout for dummy variable model, Example 5.5

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.453[a]	.205	.146	168.948

a. Predictors: (Constant), X2, X1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	198772.5	2	99386.233	3.482	.045[a]
	Residual	770670.9	27	28543.367		
	Total	969443.4	29			

a. Predictors: (Constant), X2, X1

b. Dependent Variable: COST

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	279.600	53.426		5.233	.000	169.979	389.221
	X1	80.300	75.556	.211	1.063	.297	-74.728	235.328
	X2	198.200	75.556	.520	2.623	.014	43.172	353.228

a. Dependent Variable: COST

Two-way ANOVA

If we have two categorial independent variables, e.g. the first one having 3 levels and the second one having 2 levels, then the first-order model (main effect model) is

$$Y = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{\text{main effect terms}} + \underbrace{\beta_3 x_3}_{\text{main effect term}} + \varepsilon, \tag{2}$$

where x_1 and x_2 are the dummy variables for the first (3-level) independent variable, and x_3 is the dummy variable for the second (2-level) independent variable.

The null hypothesis that $\beta_1 = \beta_2 = 0$ means that the first independent variable has no effect on the mean of Y, whilst the null hypothesis that $\beta_3 = 0$ means that the second independent variable has no effect.

This situation arises when we have two factors that may affect the mean, e.g. we have a factor called, say, "Group" and another factor called "Class", which may affect the mean of the response. The data will be tabulated in the format as the following table.

Table: The format of the data in two-way ANOVA

		The factor "Class"			
		Class C_1	Class C_2	$\cdots$	Class C_J
	Group G_1				
The factor "Group"	$\vdots$				
	Group G_I				

We may ask whether in different groups the means are the same, i.e. the mean of each row is the same as the means of the other rows. To answer this question, we test whether the parameters of the $I - 1$ dummies corresponding to the categorical variable "Group" are all zero or not.

We may also ask whether in different classes the means are the same, i.e. the mean of each column is the same as the means of the other columns. To answer, we test whether the parameters of the $J - 1$ dummies corresponding to the categorical variable "Class" are all zero or not.

These two tests are applied to model (2), which is called the *two-way ANOVA* model. The regression model for the ANOVA problem we encountered in MATH2206 is called the one-way ANOVA model.

When we have one-way ANOVA and two-way ANOVA, it is straightforward to generalise to k-way ANOVA model if we have k categorical variables.

The idea is to decompose the total sum of squares into $k + 1$ terms, corresponding to the k terms of the sum of squares of individual categorical variables and one more term of error sum of squares; then we can test whether each categorical variable has a sum of square that is significantly different from the error sum of squares or not by a partial F-test.

Therefore, we will carry out k partial F-tests, one for each categorical variable, in such a k-way ANOVA model. Each partial F-test is testing whether the coefficients of all dummy variables of a categorical variable, in the full model having the dummy variables of all these k categorical variables, are all zero or not.

Though we still have to carry out the k partial F-tests, this approach is different from (and more powerful than) applying the one-way ANOVA analysis to the same data k times, because the one-way ANOVA model has only one categorical variable (i.e. the regression model containing only the dummy variables of this categorical variable) and so the error sum of squares in the one-way ANOVA actually includes the variation not only from the errors but also from the missing variables, resulting in an over-estimate of the error variance σ^2.

Note that the k-way ANOVA model does not contain any continuous variables, and also note that when we talk about ANOVA of a regression model (which can contain categorical and continuous variables), we are talking the F-test of the null hypothesis that all parameters, except the intercept, are zero, while when we talk about k-way ANOVA analysis, we are talking about the partial F-test procedure applied to the parameters of the dummy variables of each categorical variable, individually, in the k-way ANOVA model. Nevertheless, these are just matters of terminology.

Now, if we add the interaction terms to the model given in (2), we have the model of 2-way ANOVA with interaction:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \underbrace{\beta_4 x_1 x_3 + \beta_5 x_2 x_3}_{\text{interaction terms}} + \varepsilon.$$

The interaction terms will involve all possible two-way cross-products between each of the two dummy variables (x_1 and x_2) for the first independent variable and the one dummy (x_3) for the second independent variable. (The product of two dummies is still a dummy and so is still of first-order.)

In general, we have

$$\begin{aligned}&\text{\# interaction terms}\\ &= \text{(\# main effect terms of the first independent variable)}\\ &\times \text{(\# main effect terms of the second independent variable).}\end{aligned}$$

Hence, if the two independent variables have I levels and J levels respectively, there are $(I-1)(J-1)$ interaction terms, plus $(I-1)+(J-1)$ main effect terms, plus 1 overall mean, giving us $I \times J$ parameters for the $I \times J$ different combinations of the levels of the two independent variables.

The interaction model will give a perfect fit for all cell-averages, where a cell-average is the sample mean of the data at a particular combination of the levels of the two independent variables.

Thus, if at each combination of the levels of the two independent variables, there is no replicate (i.e. only one observation in each cell), the interaction model will give us a perfect fit for all observations i.e. the model explains 100% the variation in Y.

However, random errors are really present, but the perfect fit model does not have the error term and hence should not be considered a good model.

Usual testing strategy in two-way ANOVA:

1. Test for interaction.
   If significant (usually $p < 0.05$), stop testing, interpret all effects.
2. Else, test for row and column effects (separately).

Table 5.7 The six combinations of fuel type and diesel engine brand

		Brand	
		B_1	B_2
FUEL TYPE	F_1	μ_{11}	μ_{12}
	F_2	μ_{21}	μ_{22}
	F_3	μ_{31}	μ_{32}

Table 5.8 Performance data for combinations of fuel type and diesel engine brand

		Brand	
		B_1	B_2
FUEL TYPE	F_1	65 73 68	36
	F_2	78 82	50 43
	F_3	48 46	61 62

Main Effects Model with Two Qualitative Independent Variables, One at Three Levels (F_1, F_2, F_3) and the Other at Two Levels (B_1, B_2)

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\substack{\text{Main effect} \\ \text{terms for } F}} + \overbrace{\beta_3 x_3}^{\substack{\text{Main effect} \\ \text{term for } B}}$$

where

$$x_1 = \begin{cases} 1 & \text{if } F_2 \\ 0 & \text{if not} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{if } F_3 \\ 0 & \text{if not} \end{cases} \qquad (F_1 \text{ is base level})$$

$$x_3 = \begin{cases} 1 & \text{if } B_2 \\ 0 & \text{if } B_1 \quad \text{(base level)} \end{cases}$$

Interpretation of Model Parameters

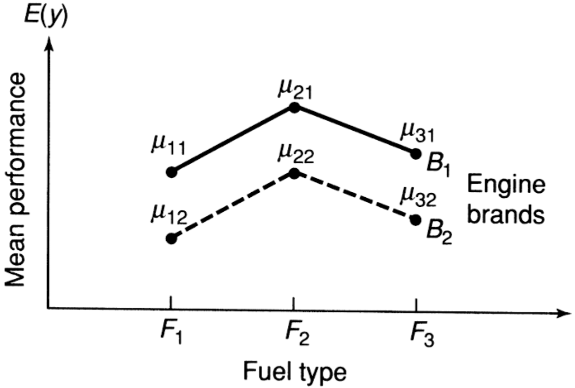
$$\beta_0 = \mu_{11} \text{ (Mean of the combination of base levels)}$$

$$\beta_1 = \mu_{2j} - \mu_{1j}, \text{ for any level } B_j (j = 1, 2)$$

$$\beta_2 = \mu_{3j} - \mu_{1j}, \text{ for any level } B_j (j = 1, 2)$$

$$\beta_3 = \mu_{i2} - \mu_{i1}, \text{ for any level } F_i (i = 1, 2, 3)$$

Figure 5.20 Main effects model: Mean response as a function of *F* and *B* when *F* and *B* affect *E*(*y*) independently



Interaction Model with Two Qualitative Independent Variables, One at Three Levels (F_1, F_2, F_3) and the Other at Two Levels (B_1, B_2)

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\text{Main effect terms for } F} + \overbrace{\beta_3 x_3}^{\text{Main effect term for } B} + \overbrace{\beta_4 x_1 x_3 + \beta_5 x_2 x_3}^{\text{Interaction terms}}$$

where the dummy variables x_1, x_2, and x_3 are defined in the same way as for the main effects model.

Interpretation of Model Parameters

$\beta_0 = \mu_{11}$ (Mean of the combination of base levels)

$\beta_1 = \mu_{21} - \mu_{11}$ (i.e., for base level B_1 only)

$\beta_2 = \mu_{31} - \mu_{11}$ (i.e., for base level B_1 only)

$\beta_3 = \mu_{12} - \mu_{11}$ (i.e., for base level F_1 only)

$\beta_4 = (\mu_{22} - \mu_{12}) - (\mu_{21} - \mu_{11})$

$\beta_5 = (\mu_{32} - \mu_{12}) - (\mu_{31} - \mu_{11})$

Figure 5.21 Interaction model: Mean response as a function of *F* and *B* when *F* and *B* interact to affect *E*(*y*)

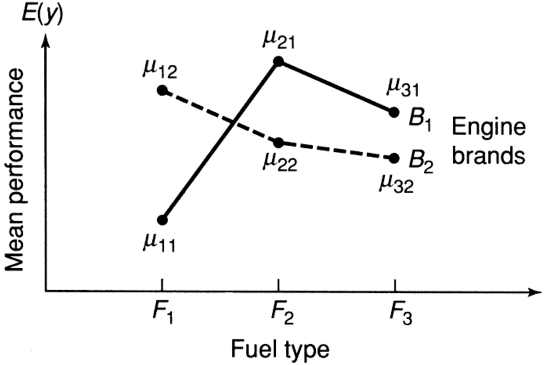


Figure 5.22 SAS printout for main effects model, Example 5.10

Dependent Variable: PERFORM

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	858.25758	286.08586	1.51	0.2838
Error	8	1512.40909	189.05114		
Corrected Total	11	2370.66667			

Root MSE	13.74959	R-Square	0.3620
Dependent Mean	59.33333	Adj R-Sq	0.1228
Coeff Var	23.17346		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	64.45455	7.18049	8.98	<.0001
X1	1	6.70455	9.94093	0.67	0.5190
X2	1	-2.29545	9.94093	-0.23	0.8232
X3	1	-15.81818	8.29131	-1.91	0.0928

Output Statistics

Obs	FUELBRND	Dep Var PERFORM	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	F1B1	65.0000	64.4545	7.1805	47.8963	81.0128	0.5455
2	F1B1	73.0000	64.4545	7.1805	47.8963	81.0128	8.5455
3	F1B1	68.0000	64.4545	7.1805	47.8963	81.0128	3.5455
4	F1B2	36.0000	48.6364	9.2700	27.2598	70.0130	-12.6364
5	F2B1	78.0000	71.1591	8.0280	52.6464	89.6718	6.8409
6	F2B1	82.0000	71.1591	8.0280	52.6464	89.6718	10.8409
7	F2B2	50.0000	55.3409	8.0280	36.8282	73.8536	-5.3409
8	F2B2	43.0000	55.3409	8.0280	36.8282	73.8536	-12.3409
9	F3B1	48.0000	62.1591	8.0280	43.6464	80.6718	-14.1591
10	F3B1	46.0000	62.1591	8.0280	43.6464	80.6718	-16.1591
11	F3B2	61.0000	46.3409	8.0280	27.8282	64.8536	14.6591
12	F3B2	62.0000	46.3409	8.0280	27.8282	64.8536	15.6591

Sum of Residuals	0
Sum of Squared Residuals	1512.40909
Predicted Residual SS (PRESS)	3615.37520

Figure 5.23 SAS printout for interaction model, Example 5.10

Dependent Variable: PERFORM

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2303.00000	460.60000	40.84	0.0001
Error	6	67.66667	11.27778		
Corrected Total	11	2370.66667			

Root MSE	3.35824	R-Square	0.9715
Dependent Mean	59.33333	Adj R-Sq	0.9477
Coeff Var	5.65996		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	68.66667	1.93888	35.42	<.0001
X1	1	11.33333	3.06564	3.70	0.0101
X2	1	-21.66667	3.06564	-7.07	0.0004
X3	1	-32.66667	3.87776	-8.42	0.0002
X1X3	1	-0.83333	5.12980	-0.16	0.8763
X2X3	1	47.16667	5.12980	9.19	<.0001

Output Statistics

Obs	FUELBRND	Dep Var PERFORM	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	F1B1	65.0000	68.6667	1.9389	63.9224	73.4109	-3.6667
2	F1B1	73.0000	68.6667	1.9389	63.9224	73.4109	4.3333
3	F1B1	68.0000	68.6667	1.9389	63.9224	73.4109	-0.6667
4	F1B2	36.0000	36.0000	3.3582	27.7827	44.2173	-7.11E-15
5	F2B1	78.0000	80.0000	2.3746	74.1895	85.8105	-2.0000
6	F2B1	82.0000	80.0000	2.3746	74.1895	85.8105	2.0000
7	F2B2	50.0000	46.5000	2.3746	40.6895	52.3105	3.5000
8	F2B2	43.0000	46.5000	2.3746	40.6895	52.3105	-3.5000
9	F3B1	48.0000	47.0000	2.3746	41.1895	52.8105	1.0000
10	F3B1	46.0000	47.0000	2.3746	41.1895	52.8105	-1.0000
11	F3B2	61.0000	61.5000	2.3746	55.6895	67.3105	-0.5000
12	F3B2	62.0000	61.5000	2.3746	55.6895	67.3105	0.5000

Sum of Residuals	0
Sum of Squared Residuals	67.66667
Predicted Residual SS (PRESS)	213.50000

Figure 5.25 SAS printout for nested model *F*-test of interaction

Test INTERACT Results for Dependent Variable PERFORM

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	722.37121	64.05	<.0001
Denominator	6	11.27778		

Models with both quantitative and qualitative independent variables

Consider first a model with one continuous independent variable x_1 and a three-level categorical independent variable, which requires two dummy variables x_2 and x_3. The quadratic model without interaction is

$$Y = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_1^2}_{\substack{\text{continuous,} \\ \text{second-order,} \\ \text{main effects}}} + \underbrace{\beta_3 x_2 + \beta_4 x_3}_{\substack{\text{dummies,} \\ \text{main effects}}} + \varepsilon,$$

which will give us three parallel curves (more precisely, three parallel parabolas), each corresponds to one different level of the categorical independent variable.

If we include interactions, then we have the complete second-order model:

$$\begin{aligned} Y &= \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_1^2}_{\text{main effects}} + \underbrace{\beta_3 x_2 + \beta_4 x_3}_{\text{main effects}} \\ &\quad + \underbrace{\beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1^2 x_2 + \beta_8 x_1^2 x_3}_{\text{continuous-categorical interaction terms}} + \varepsilon, \end{aligned} \tag{3}$$

which is still a second-order model, because the dummy variables in $x_1^2 x_2$ and $x_1^2 x_3$ are not making any contribution to the order and so these two terms are still second-order.

We will get three different parabolas for three different levels of the categorical predictor. This requires nine parameters (3 parameters per parabola $\times$ 3 parabolas).

The estimates of the parameters are the same as if we split the data into three groups, corresponding to the three different levels, and then fit a quadratic regression model to each group.

Fitting one model with all interaction terms and fitting three separate quadratic regression models individually give the same three fitted parabolas.

Thus, why don't we write three separate models? The reason are:

1. In the estimation of the parameters of the curve at level 1 (β_1 and β_2) we use information of the data from levels 2 and 3 when fitting model (3), whereas we would not use any information of the data from levels 2 and 3 when fitting three separate quadratic regression models.
2. If we write three models, then we will have three different variances σ_1^2, σ_2^2 and σ_3^2 for the error terms. If we assume the variances are the same and equal to σ^2, then using one single model allows us to obtain a pooled estimate of σ^2 (i.e. an estimate from the pooled data).
3. Moreover, using one single model allows us to use partial F-tests to test whether the parabolas are parallel or to test any nested models.

Suppose now we have one continuous independent variable x_1 and two categorical independent variables, each of which has two levels. Then the complete second-order model is

$$
\begin{aligned}
Y &= \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_1^2}_{\substack{\text{continuous,}\\ \text{second-order,}\\ \text{main effects}}} + \underbrace{\beta_3 x_2 + \beta_4 x_3}_{\substack{\text{dummies,}\\ \text{main effects}}} + \underbrace{\beta_5 x_2 x_3}_{\text{categorical-categorical interaction term}} \\
&+ \underbrace{\beta_6 x_1 x_2 + \beta_7 x_1 x_3 + \beta_8 x_1 x_2 x_3}_{\text{(first-order) continuous-categorical interaction terms}} \\
&+ \underbrace{\beta_9 x_1^2 x_2 + \beta_{10} x_1^2 x_3 + \beta_{11} x_1^2 x_2 x_3}_{\text{(second-order) continuous-categorical interaction terms}} + \varepsilon,
\end{aligned}
$$

which requires twelve parameters (3 parameters per parabola $\times$ 4 parabolas).

How many parameters do we need for, say, two continuous independent variables and two categorical independent variables, one of which has two and the other has three levels? A paraboloid (or a saddle-shape) surface requires 6 parameters and there are 2×3 combinations of the two categorical variables. It will be better to count in a more systematic way.

How many?	for
1	intercept β_0,
5	all first- and second-order terms, including interaction, of the two continuous independent variables, i.e., x_1, x_2, x_1^2, x_2^2, x_1x_2,
5	one dummy x_3 for the first, two dummies x_4 and x_5 for the second categorical independent variables, and two interaction terms x_3x_4 and x_3x_5,
5×5	all (first- and second-order) continuous-categorical interaction terms.
36	the full model in total.

We can see how important it is to carefully select the independent variables to be considered, because even for such a simple model with just four independent variables, we need 36 parameters!

Chapter 6 will tell us how to perform variable screening in order to choose more important variables to be included in the model building process.

Model Validation

Models that fit the sample data well may not be a successful model for prediction of Y when applied to new data.

For this reason, it is important to assess the *validity* (how successful it will be, when applied to new or future data) of the regression model in addition to its *adequacy* (how adequate the model is, when used to fit the sample data) before using it in practice.

Five ways to assess its validity are as follows.

(i) *Examining the predicted values*: The predicted values $\hat{Y}$ can help to identify an invalid model. Nonsensical or unreasonable predicted values may indicate that the form of the model is incorrect or that the coefficients are poorly estimated.

(ii) *Examining the estimated model parameters*: Prior information on the relative size and sign of the model parameters could be used as a check on the estimated coefficients.

(iii) *Collecting new data for prediction*: One of the most effective ways is to use the model to predict Y for a new sample and then compare them with the new observations. Suppose the new sample of size m is $\{Y_{n+1}, \ldots, Y_{n+m}\}$, we can consider the following measures of model validity:

(a)
$$R^2_{\text{prediction}} := 1 - \frac{\sum_{i=n+1}^{n+m}(Y_i - \hat{Y}_i)^2}{\sum_{i=n+1}^{n+m}(Y_i - \overline{Y})^2},$$

where $\overline{Y}$ is the sample mean of the original data (alternatively, the sample mean of the new data may be used), and $\hat{Y}_i$ is the predicted value using the fitted model.

If $R^2_{\text{prediction}}$ compares favourably to R^2, then the model seems trustworthy for prediction. If there is a substantial drop, then we should be cautious.

(b)
$$MSE_{\text{prediction}} := \frac{\sum_{i=n+1}^{n+m}(Y_i - \hat{Y}_i)^2}{m-k-1},$$

which should be comparable to the MSE of the least squares fit.

For either one, the new data set should be large enough to reliably assess the model's prediction performance and it has been suggested that at least 15–20 new observations are needed.

(iv) *Cross-validation (data-splitting)*: If no new data are available, the original data can be split into two parts, with one part used to estimate and the other to calculate $R^2_{\text{prediction}}$ and $MSE_{\text{prediction}}$ to assess the fitted model's predictivity ability. Random splits are usually applied in cases where there is no logical basis for dividing the data. In this case, we should have at least $n = 2k + 25$ observations for a model with k independent variables.

(v) *Jackknifing*: The jackknife method involves leaving each observation out of the data set, one at a time, and calculating the difference $Y_i - \hat{Y}_{(i)}$ for all observations in the data set, where $\hat{Y}_{(i)}$ denotes the predicted value for the i^{th} observation obtained when the regression model is fitted <u>without</u> the data point for Y_i, so that $\hat{Y}_{(i)}$ and Y_i are independent. We can then calculate

$$PRESS = \text{prediction sum of squares} := \sum_{i=1}^{n} (Y_i - \hat{Y}_{(i)})^2,$$

$$R^2_{\text{jackknife}} := 1 - \frac{PRESS}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2},$$

$$MSE_{\text{jackknife}} := \frac{PRESS}{n - k - 1}.$$

Since least squares estimation (LSE) will minimize $SSE := \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$ and so the least squares fit $\hat{Y}_i$ should be closer to Y than the jackknife prediction $\hat{Y}_{(i)}$ should be, suggesting that in general

$$SSE < PRESS,$$

which implies that $R^2_{\text{jackknife}} < R^2$ and $MSE_{\text{jackknife}} > MSE$. However, the model parameters used in getting $\hat{Y}_{(i)}$ depend on i and hence are not fixed; that is to say, this argument could not immediately lead to these inequalities.

However, these inequalities are true in general. (Later in this course we will discuss the matrix $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Denote its i^{th} diagonal element by h_{ii}, which is called the i^{th} leverage. It can be shown that $0 \le h_{ii} \le 1$ and $PRESS = \sum_i \{e_i/(1-h_{ii})\}^2 > \sum_i e_i^2 = SSE$. This also shows that we need not fit the regression model repeatedly to get all $\hat{Y}_{(i)}$ for the calculation of $PRESS$.)

When $R^2_{\text{jackknife}}$ (or $MSE_{\text{jackknife}}$) is reasonably close to R^2 (or MSE, respectively), the validity of the model is good.