Contents

Contents **1**

5 Principles of Model Building **2**

5.1 Introduction: Why Model Building Is Important 2

5.2 Models with a Single Quantitative Independent Variable 2

5.3 Models with Two Qualitative Independent Variables 8

5.4 Models with Three or More Qualitative Independent Variables 14

5.5 Models with Both Quantitative and Qualitative Independent Variables 15

Chapter 5

Principles of Model Building

In this chapter, we explain the importance of the deterministic portion of a linear model. We also present models with only numerical variables, only categorical variables, or both types of variables. Finally, we explain some basic procedures for building good linear models.

5.1 Introduction: Why Model Building Is Important

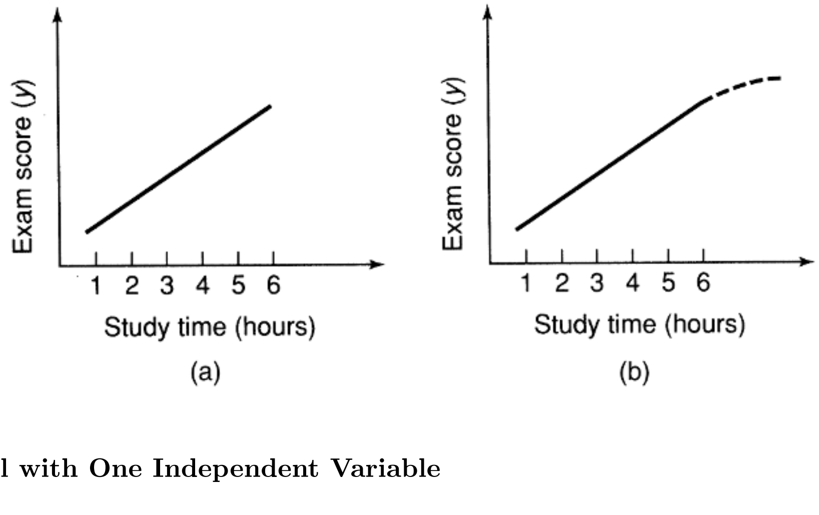
As we mentioned before, one of the first steps in the construction of a regression model is to hypothesize the form of the deterministic portion of the probabilistic model.

Model Building

Writing a model that will provide a good fit to a set of data and that will give good estimates of the mean value of y and good predictions of future values of y for given values of the independent variables.

5.2 Models with a Single Quantitative Independent Variable

The following plots show the relationship between exam score, y, and the amount of study time, x for different ranges of x-values. Explain possible models.



A p^{th}-Order Polynomial with One Independent Variable

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + ... + \beta_p x^p$$

where p is an integer and β_0, β_1 , . . . , β_p are unknown parameters that must be estimated.

First-Order (Straight-Line) Model with One Independent Variable

When $p = 1$:

$$E(y) = \beta_0 + \beta_1 x$$

Interpretation of model parameters

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

β_1: Slope of the line; the change in $E(y)$ for a 1-unit increase in x

A Second-Order (Quadratic) Model with One Independent Variable

When $p = 2$:

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

Interpretation of model parameters

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

β_1: Shift parameter; changing the value of β_1 shifts the parabola to the right or left

β_2: Rate of curvature

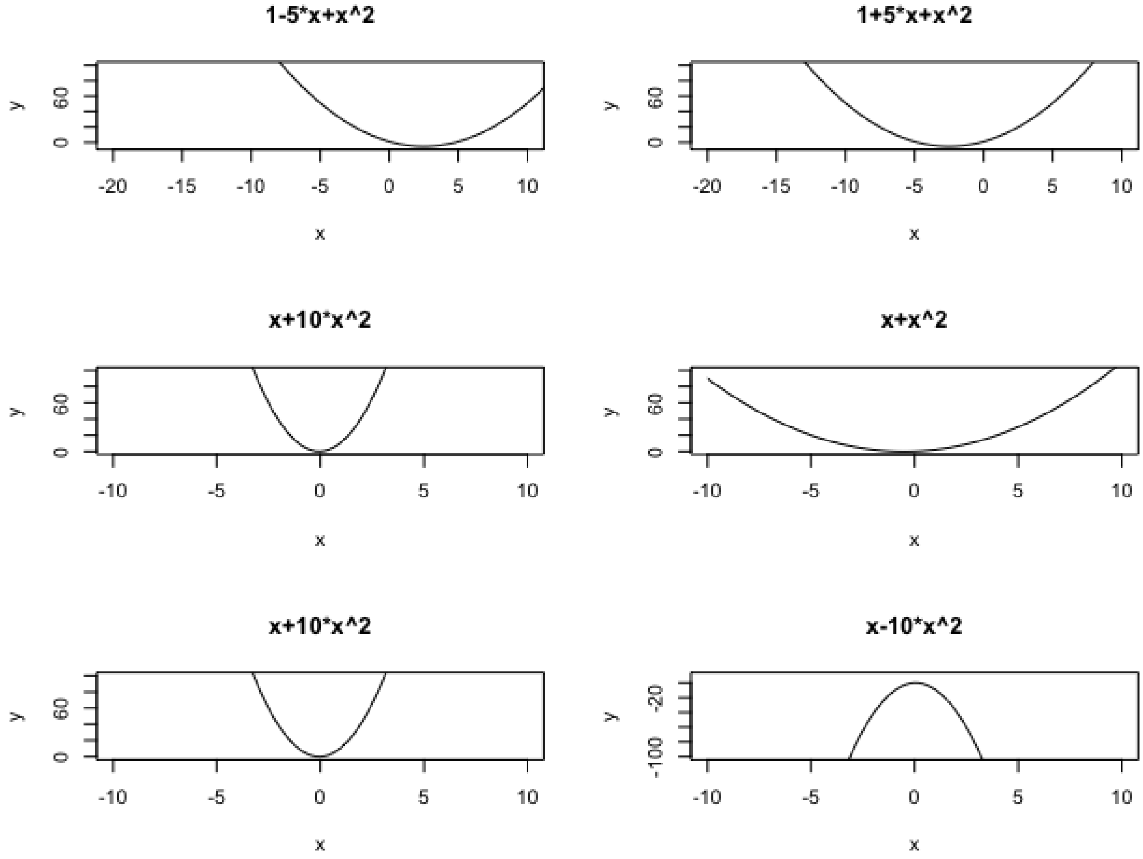


Figure 5.1: Graphs for different second-order polynomial models

Consider the following graph:

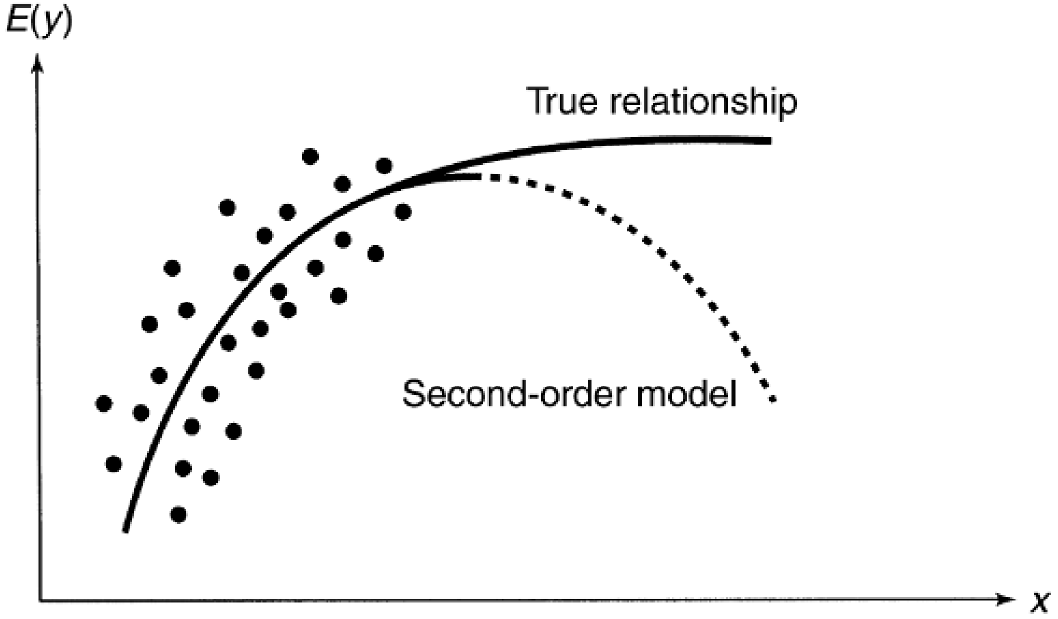


Figure 5.2: Example of the use of a quadratic model

The model is valid only over the range of x-values that were used to fit the model.

The Order of the Polynomial Model

- Prior information about the relationship between $E(y)$ and x
- Constructing a scatterplot of the data points
- A p^{th}-order polynomial for $(p-1)$ peaks
- A straight line might be used if the rate of curvature of the response curve is very small over the desired range of x

Third-Order Model with One Independent Variable

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

Interpretation of model parameters

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

β_1: Shift parameter (shifts the polynomial right or left on the x-axis)

β_2: Rate of curvature

β_3: The magnitude of β_3 controls the rate of reversal of curvature for the polynomial

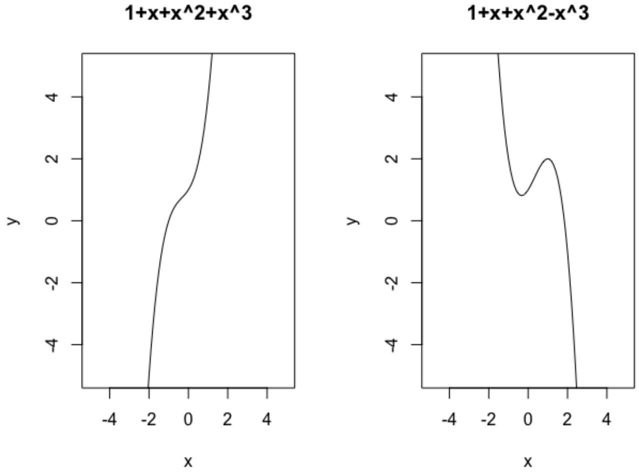
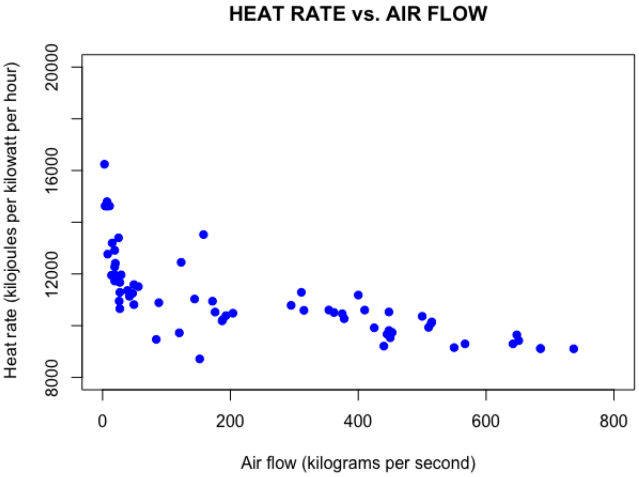


Figure 5.3: Graphs of two third-order polynomial models

Example 5.2.1 *Refer to the Cooling method for gas turbines data set.*

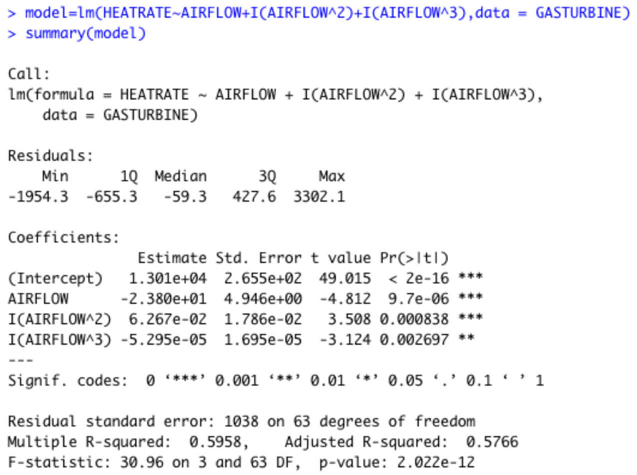
a) Construct a scatterplot relating heat rate to airflow. What type of model is suggested by the plot?



b) Fit the third-order model to the data for the heat rate (y) using AIRFLOW. Is there evidence that the cubic term, $\beta_3 x_3$, contributes information for the prediction of heat rate? Test at $\alpha = .05$.

Third-order model:

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$



```
> model=lm(HEATRATE~AIRFLOW+I(AIRFLOW^2)+I(AIRFLOW^3),data = GASTURBINE)
> summary(model)

Call:
lm(formula = HEATRATE ~ AIRFLOW + I(AIRFLOW^2) + I(AIRFLOW^3), 
    data = GASTURBINE)

Residuals:
    Min      1Q  Median      3Q     Max 
-1954.3  -655.3   -59.3   427.6  3302.1 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.301e+04  2.655e+02  49.015  < 2e-16 ***
AIRFLOW      -2.380e+01  4.946e+00  -4.812  9.7e-06 ***
I(AIRFLOW^2)  6.267e-02  1.786e-02   3.508 0.000838 ***
I(AIRFLOW^3) -5.295e-05  1.695e-05  -3.124 0.002697 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1038 on 63 degrees of freedom
Multiple R-squared:  0.5958,	Adjusted R-squared:  0.5766 
F-statistic: 30.96 on 3 and 63 DF,  p-value: 2.022e-12
```

$H_0 : \beta_3 = 0$

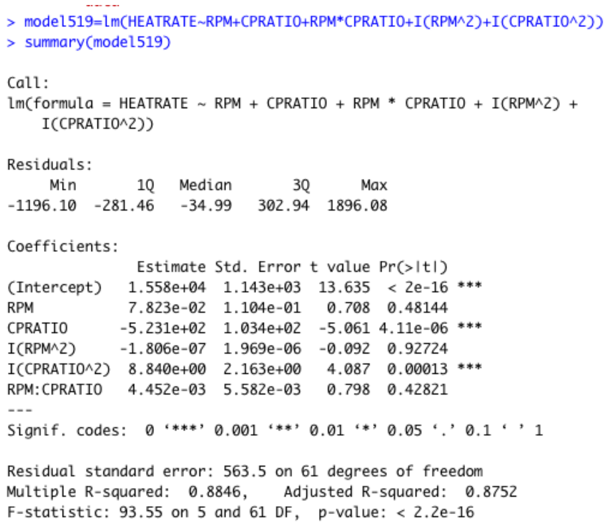
$H_a : \beta_3 \neq 0$

Since the p-value$= 0.002697 < 0.05$ for the above test, there is sufficient evidence of a third-order relationship between heat rate and air flow.

c) Write a complete second-order model for heat rate (kilojoules per kilowatt per hour)(y) as a function of cycle speed (revolutions per minute) and cycle pressure ratio. Recall that the data are saved in the GASTURBINE file.

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

d) Fit the model to the data and give the least squares prediction equation.



```
> model519=lm(HEATRATE~RPM+CPRATIO+RPM*CPRATIO+I(RPM^2)+I(CPRATIO^2))
> summary(model519)

Call:
lm(formula = HEATRATE ~ RPM + CPRATIO + RPM * CPRATIO + I(RPM^2) + 
    I(CPRATIO^2))

Residuals:
     Min       1Q   Median       3Q      Max 
-1196.10  -281.46   -34.99   302.94  1896.08 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.558e+04  1.143e+03  13.635  < 2e-16 ***
RPM           7.823e-02  1.104e-01   0.708  0.48144    
CPRATIO      -5.231e+02  1.034e+02  -5.061 4.11e-06 ***
I(RPM^2)     -1.806e-07  1.969e-06  -0.092  0.92724    
I(CPRATIO^2)  8.840e+00  2.163e+00   4.087  0.00013 ***
RPM:CPRATIO   4.452e-03  5.582e-03   0.798  0.42821    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 563.5 on 61 degrees of freedom
Multiple R-squared:  0.8846,	Adjusted R-squared:  0.8752 
F-statistic: 93.55 on 5 and 61 DF,  p-value: < 2.2e-16
```

The least squares prediction equation according to the Rstudio output is:

$$\hat{y} = 15580 + 0.07823x_1 - 523.1x_2 - 0.00000018x_1^2 + 8.84x_2^2 + 0.004452x_1x_2$$

When $x_1 =$ RPM, $x_2 =$ CPRATIO

e) Conduct a global F-test for overall model adequacy.

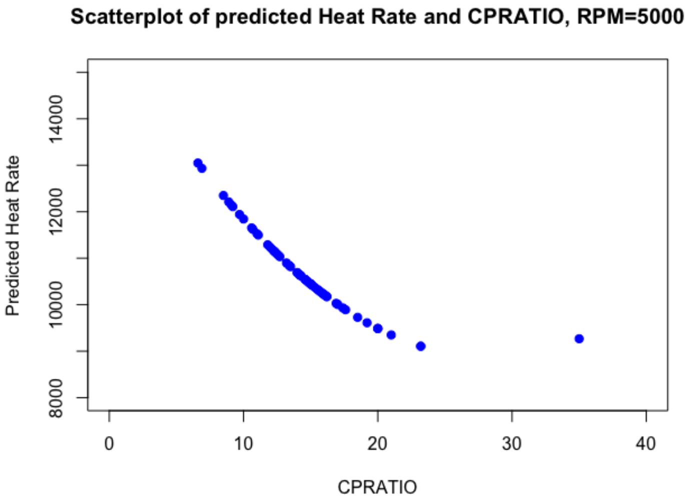
$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

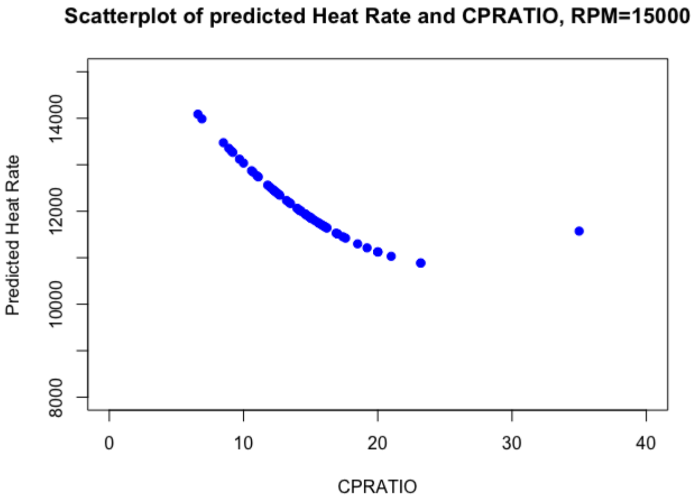
H_a : at least one $\beta_i \neq 0$

$F - statistic : 93.55$ on 5 and 61 DF, p-value: $< 2.2e - 16$

There is sufficient evidence to indicate that the model is statistically useful for predicting heat rate.

e) Based on the prediction equation, graph the relationship between heat rate and cycle pressure ratio when cycle speed is held constant at 5,000 rpm. Repeat part d when cycle speed is held constant at 15,000 rpm. Compare the two graphs, parts d and e. What do you observe?





5.3 Models with Two Qualitative Independent Variables

Suppose we want to write a model for the mean performance, $E(y)$, of a diesel engine as a function of type of fuel and engine brand. Further suppose there are three fuel types available: a petroleum-based fuel (P), a coal-based fuel (C), and a blended fuel (B) and two brands; B_1 and B_2.

Define the symbols for the three fuel types F_1, F_2, F_3, and we will let B_1 and B_2 represent the two brands.

The six population means of performance measurements (measurements of y) are symbolically represented by the six cells in the following two-way table.

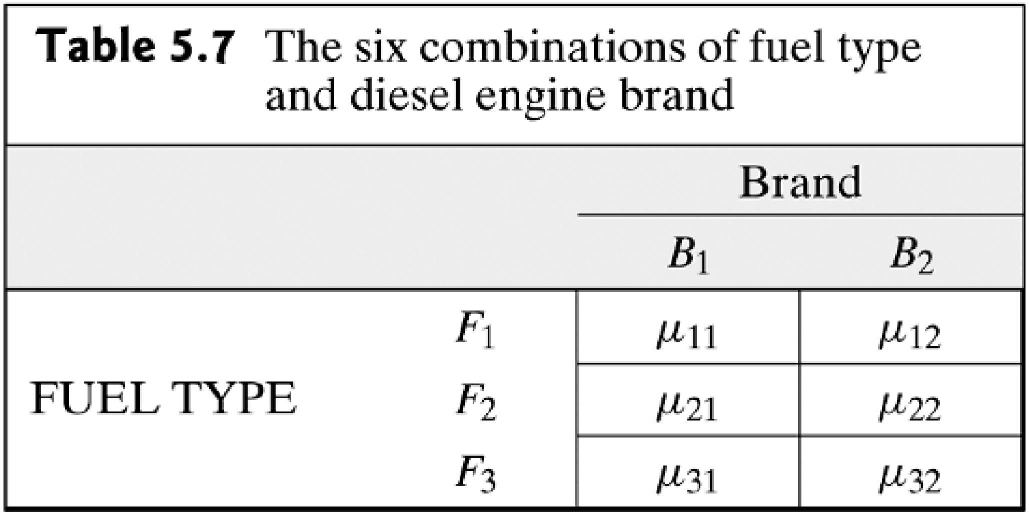


Table 5.7 The six combinations of fuel type and diesel engine brand

		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}

The Simplest Model with Two Qualitative Independent Variables, Main Effects Model

In this model, the two qualitative variables affect the response independently of each other. Changing the level of one qualitative variable will have the same effect on $E(y)$ for any level of the second qualitative variable. In other words, the effect of one qualitative variable on $E(y)$ is independent (in a mathematical sense) of the level of the second qualitative variable.

We also need $2 + 1 = 3$ dummy variables. Why?

Step 1: Adding type of fuel using two dummy variables for the three levels of fuel type.

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

$x_1 = \begin{cases} 1 & \text{if fuel type} \quad F_2 \quad \text{was used} \\ 0 & \text{if not} \end{cases}$ $x_2 = \begin{cases} 1 & \text{if fuel type} \quad F_3 \quad \text{was used} \\ 0 & \text{if not} \end{cases}$ Base Level $=$ fuel type F_1

Step 2: Adding type of brand using one dummy variable for the two levels of brand.

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Where:

Main effect terms for Fuel Type: $\beta_1 x_1 + \beta_2 x_2$

Main effect term for Brand: $\beta_3 x_3$

$$x_3 = \begin{cases} 1 & \text{if engine brand} \quad B_2 \quad \text{was used} \\ 0 & \text{if engine brand} \quad B_1 \quad \text{was used} \end{cases}$$

a) Give the values of x_1, x_2, and x_3 and the model for the mean performance, $E(y)$, when using fuel type F_1 in engine brand B_1.

$x_1 = x_2 = x_3 = 0 \longrightarrow E(y) = \mu_{11} = \beta_0$

b) Give the values of x_1, x_2, and x_3 and the model for the mean performance, $E(y)$, when using fuel type F_3 in engine brand B_2.

$x_1 = 0, x_2 = 1, x_3 = 1 \longrightarrow E(y) = \mu_{32} = \beta_0 + \beta_2 + \beta_3$

The following graph shows that the two independent variables affect the mean response independently of each other.

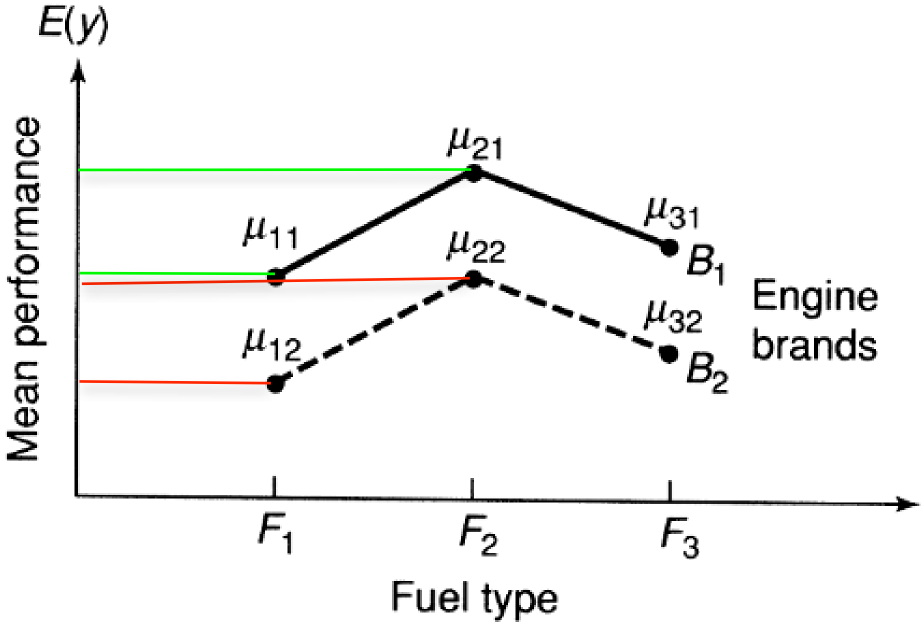


Figure 5.4: Main effects model: Mean response as a function of F and B when F and B affect $E(y)$ independently

Note: The difference in mean performance between any two fuel types (levels of F) is the same, regardless of the engine brand used. That is, the main effects model assumes that the relative effect of fuel type on performance is the same in both engine brands.

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)

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Where:

Main effect terms for Fuel Type: $\beta_1 x_1 + \beta_2 x_2$

Main effect term for Brand: $\beta_3 x_3$

Where:

$$x_1 = \begin{cases} 1 & \text{if} \quad F_2 \\ 0 & \text{if not} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{if} \quad F_3 \\ 0 & \text{if not} \end{cases} \qquad \text{Base Level} = F_1$$

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

Interpretation of Model Parameters:

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

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

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

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

The following graph shows the response function, if F and B do not affect $E(y)$ independently of each other.

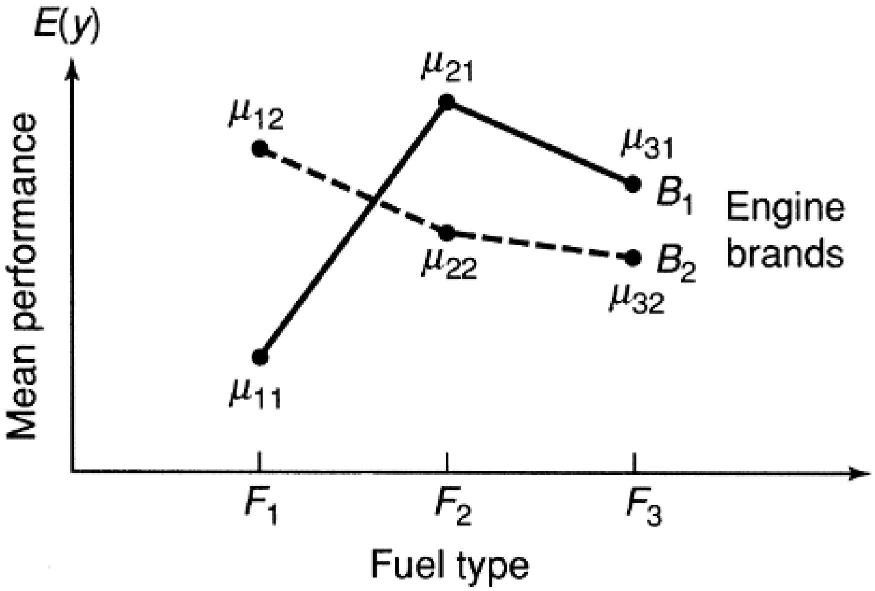


Figure 5.5: Interaction model: Mean response as a function of F and B when F and B interact to affect $E(y)$

Note: In this model we cannot study the effect of one variable on $E(y)$ without considering the level of the other. When this situation occurs, we say that the qualitative independent variables interact.

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)

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3$

Where:

Main effect terms for Fuel Type: $\beta_1 x_1 + \beta_2 x_2$

Main effect term for Brand: $\beta_3 x_3$

Interaction terms: $\beta_4 x_1 x_3 + \beta_5 x_2 x_3$

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)

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

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

Number of Interaction Terms:

If we have two indepndent variables, the number of interaction terms will equal the number of main effect terms for the one variable times the number of main effect terms for the other.

c) Give the value of $E(y)$ for the model where F and B interact to affect $E(y)$ and when fuel F_1 was used in engine B_1.

$x_1 = x_2 = x_3 = 0 \longrightarrow E(y) = \mu_{11} = \beta_0$

d) Now assume that F and B interact, and give the value for $E(y)$ when fuel F_3 is used in engine brand B_2.

$x_1 = 0, x_2 = 1, x_3 = 1 \longrightarrow E(y) = \mu_{32} = \beta_0 + \beta_2 + \beta_3 + \beta_5$

Example 5.3.1 *The performance, y (measured as mass burning rate per degree of crank angle), for the six combinations of fuel type and engine brand is shown in Table 5.8. The number of test runs per combination varies from one for levels (F_1, B_2) to three for levels (F_1, B_1). Dtat set: DIESEL*

A total of 12 test runs are sampled.

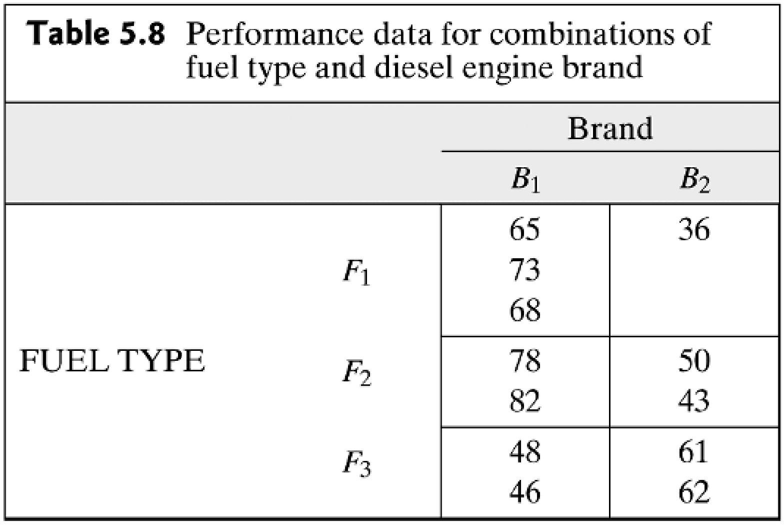


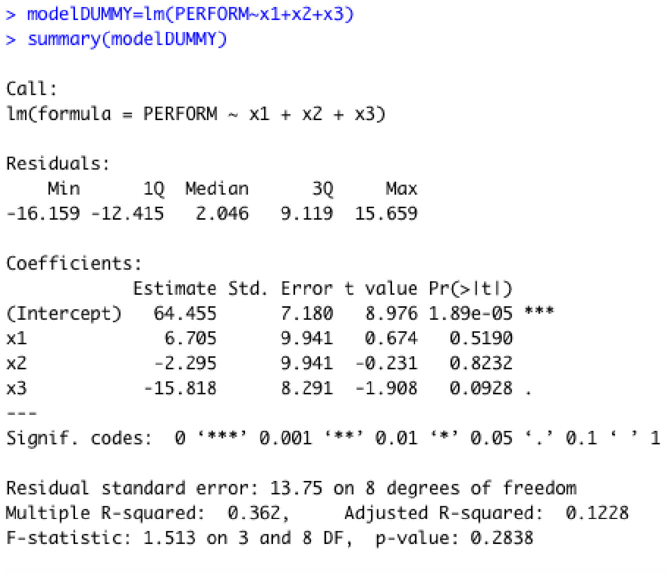
Table 5.8 Performance data for combinations of fuel type and diesel engine brand

		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

(a) Assume the interaction between F and B is negligible. Fit the model for $E(y)$ with interaction terms omitted.

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

$\hat{y} = 64.455 + 6.705x_1 - 2.295x_2 - 15.818x_3$

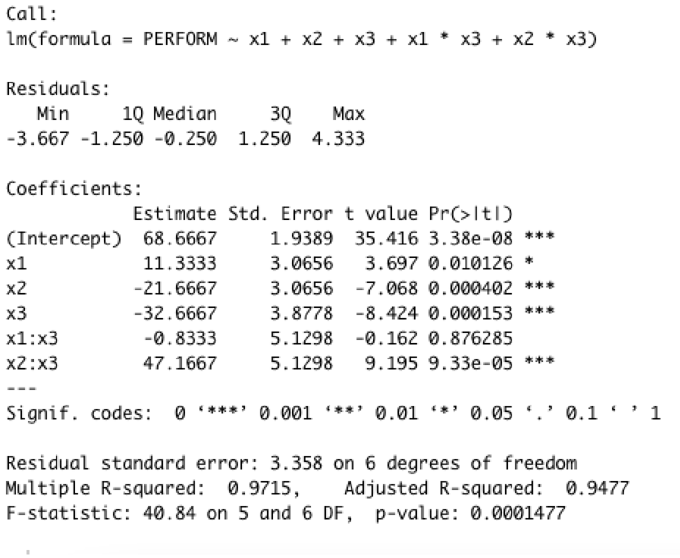


```
> modelDUMMY=lm(PERFORM~x1+x2+x3)
> summary(modelDUMMY)

Call:
lm(formula = PERFORM ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.159 -12.415   2.046   9.119  15.659 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   64.455      7.180   8.976 1.89e-05 ***
x1             6.705      9.941   0.674   0.5190    
x2            -2.295      9.941  -0.231   0.8232    
x3           -15.818      8.291  -1.908   0.0928 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.75 on 8 degrees of freedom
Multiple R-squared:  0.362,	Adjusted R-squared:  0.1228 
F-statistic: 1.513 on 3 and 8 DF,  p-value: 0.2838
```

Figure 5.6: RStudio printout for main effects model

(b) Fit the complete model for E(y) allowing for the fact that interactions might occur.

$\hat{y} = 68.6667 + 11.3333x_1 - 21.6667x_2 - 32.6667x_3 - 0.8333x_1x_3 + 47.1667x_2x_3$



```
Call:
lm(formula = PERFORM ~ x1 + x2 + x3 + x1 * x3 + x2 * x3)

Residuals:
   Min     1Q Median     3Q    Max 
-3.667 -1.250 -0.250  1.250  4.333 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  68.6667     1.9389  35.416 3.38e-08 ***
x1           11.3333     3.0656   3.697 0.010126 *  
x2          -21.6667     3.0656  -7.068 0.000402 ***
x3          -32.6667     3.8778  -8.424 0.000153 ***
x1:x3        -0.8333     5.1298  -0.162 0.876285    
x2:x3        47.1667     5.1298   9.195 9.33e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.358 on 6 degrees of freedom
Multiple R-squared:  0.9715,	Adjusted R-squared:  0.9477 
F-statistic: 40.84 on 5 and 6 DF,  p-value: 0.0001477
```

Figure 5.7: RStudio printout for the complete model

(c) Use the prediction equation for the model, part a to estimate the mean engine performance when fuel F_3 is used in brand B_2. Then calculate the sample mean for this cell in Table 5.8. Repeat for the model, part b. Explain the discrepancy between the sample mean for levels (F_3, B_2) and the estimate(s) obtained from one or both of the two prediction equations.

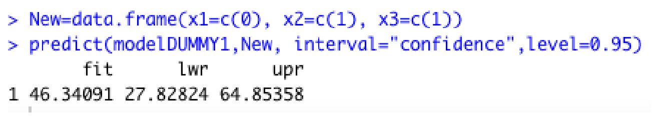
For the main effects model:

$x_1 = 0, x_2 = 1, x_3 = 1 \longrightarrow \hat{y} = 64.455 + 6.705(0) - 2.295(1) - 15.818(1) = 46.342$

The follwoing output shows the 95% CI for the true mean performance:

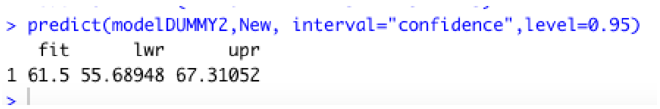
For the complete model:

$\hat{y} = 68.6667 + 11.3333(0) - 21.6667(1) - 32.6667(1) - 0.8333(0)(1) + 47.1667(1)(1) = 61.5$



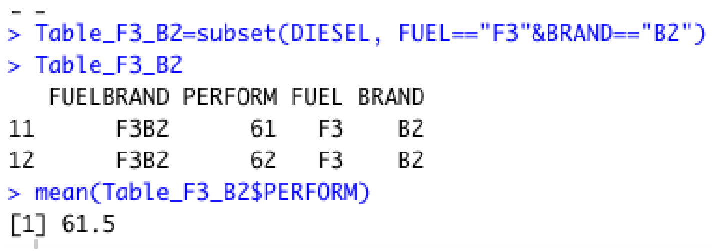
```
> New=data.frame(x1=c(0), x2=c(1), x3=c(1))
> predict(modelDUMMY1,New, interval="confidence",level=0.95)
       fit      lwr      upr
1 46.34091 27.82824 64.85358
```

The follwoing output shows the 95% CI for the true mean performance using complete model:



```
> predict(modelDUMMY2,New, interval="confidence",level=0.95)
   fit      lwr      upr
1 61.5 55.68948 67.31052
> 
```

The following output shows the observations for combinationsof fuel type of F_3 and diesel engine brand of B_2 and the sample mean for the corresponding cell (F_3, B_2), $\bar{y}_{32}$.

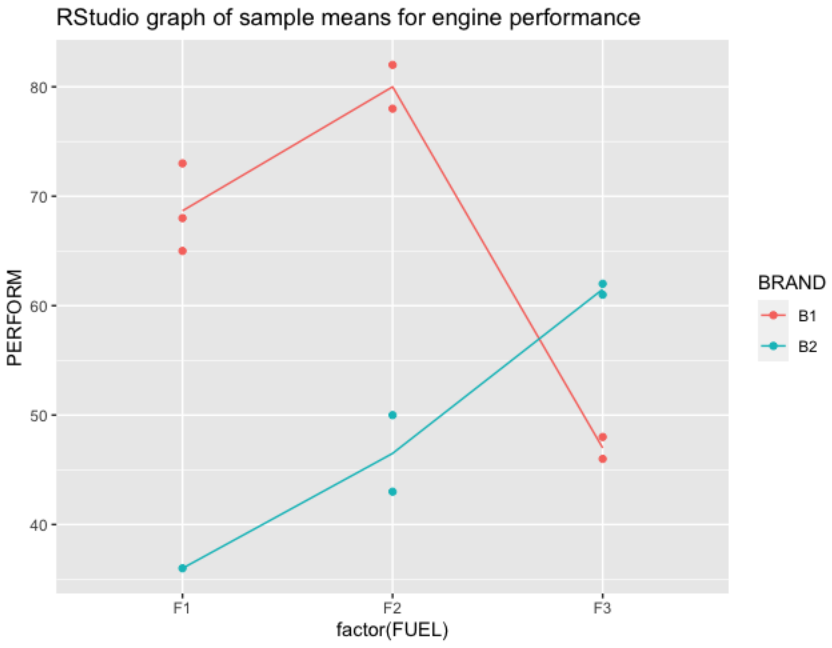


```
> Table_F3_B2=subset(DIESEL, FUEL=="F3"&BRAND=="B2")
> Table_F3_B2
   FUELBRAND PERFORM FUEL BRAND
11      F3B2      61   F3    B2
12      F3B2      62   F3    B2
> mean(Table_F3_B2$PERFORM)
[1] 61.5
```

$\bar{y}_{32}$ is precisely estimated by the complete (interaction) model. Hence, the discrepancy between the estimated mean performance using the main effects model and the sample mean for levels (F_3, B_2) is becuase the main effects model assumes the two qualitative independent variables affect $E(y)$ independently of each other. Thus, the complete model estimate for any cell mean is equal to the observed (sample) mean for that cell.

d) Do the data provide sufficient evidence to indicate that F and B interact?

Question: What do you conclude from the following plot?



In order to generalize these sample facts to the populations we perform a test:

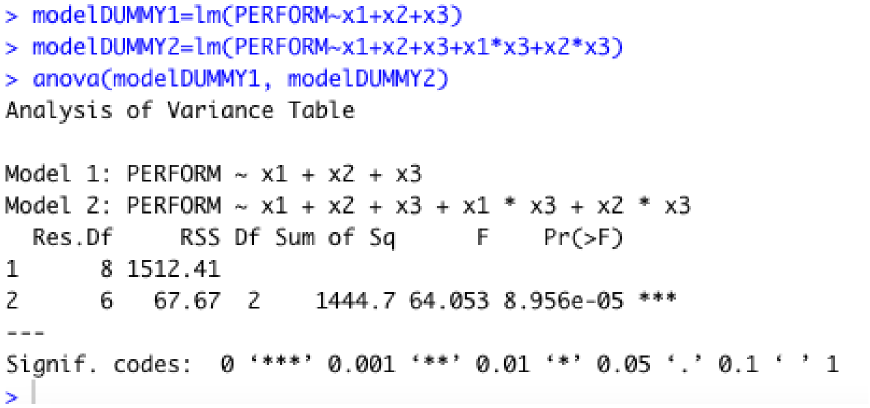
$H_0 : \beta_4 = \beta_5 = 0$

H_a : At least one of β_4 and β_5 differs from 0

Reduced model: Main Effects Model

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Complete model: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3$



```
> modelDUMMY1=lm(PERFORM~x1+x2+x3)
> modelDUMMY2=lm(PERFORM~x1+x2+x3+x1*x3+x2*x3)
> anova(modelDUMMY1, modelDUMMY2)
Analysis of Variance Table

Model 1: PERFORM ~ x1 + x2 + x3
Model 2: PERFORM ~ x1 + x2 + x3 + x1 * x3 + x2 * x3
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)    
1      8 1512.41                                  
2      6   67.67  2    1444.7 64.053 8.956e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Figure 5.8: RStudio printout for nested model F -test of interaction

The test statistic, $F = 64.053$ and p-value $= 8.956e - 05 < 0.05$.

At 5% level of significance, we have sufficient evidence to conclude that the interaction terms contribute to the prediction of y, engine performance. Equivalently, there is sufficient evidence to conclude that factors F and B do interact.

5.4 Models with Three or More Qualitative Independent Variables

Constructing models with three or more categorical variables are similar to models for two categorical independent variables, except that we must add three-way interaction terms if we have three qualitative independent variables, three-way and four-way interaction terms for four independent variables, and so on.

Pattern of the Model Relating $E(y)$ to k Qualitative Independent Variables

$$
\begin{aligned}
E(y) = \beta_0 + &\text{ Main effect terms for all independent variables} \\
&+ \text{All two-way interaction terms between pairs of independent variables} \\
&+ \text{All three-way interaction terms between different groups of three independent variables} \\
&+ \\
&. \\
&. \\
&. \\
&+ \text{All k-way interaction terms for the k independent variables}
\end{aligned}
$$

Example 5.4.1 *a) Refer to Example 5.3.1, where we modeled the performance, y , of a diesel engine as a function of fuel type (F_1, F_2, and F_3) and brand (B_1 and B_2). Now consider a third qualitative independent variable, injection system (S_1 and S_2).*

b) Refer to Example 5.3.1 and give the expression for the mean value of performance y for engines using Fuel type F_2, of Brand B_1, and with injection System S_2.

c) Suppose you want to test the hypothesis that the three qualitative independent variables discussed in part (a) do not interact, (i.e.), the hypothesis that the effect of any one of the variables on $E(y)$ is independent of the level settings of the other two variables. Formulate the appropriate test of hypothesis about the model parameters.

5.5 Models with Both Quantitative and Qualitative Independent Variables

Suppose for the example of of a diesel engine, we suppose mean performance of a diesel engine is a function of one qualitative independent variable, fuel type at levels F_1, F_2, and F_3, and one quantitative independent variable, engine speed in revolutions per minute (rpm).

First, let's assume the categorical variable has no effect on the dependent variable. That means the mean contribution to the dependent variable is the same for all the three types of fuel, but the mean performance, $E(y)$, is related to engine speed.

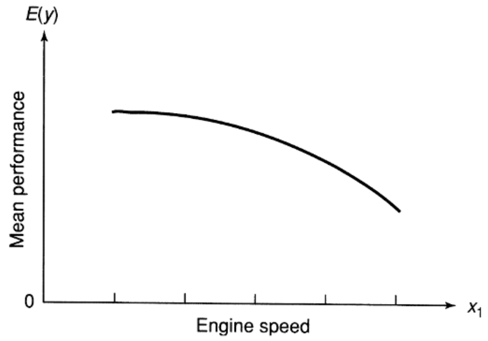


Figure 5.9: Model for $E(y)$ as a function of engine speed

If the assumption is not correct, then differences exist in mean performance for the three fuel types which would inflate the SSE associated with the fitted model and consequently would increase errors of estimation and prediction.

Second, we assume that the qualitative independent variable, fuel type, does affect mean performance, but the effect on $E(y)$ is independent of speed. (i.e. no interaction)

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3$$

$x_1 =$ Engine speed $x_2 = \begin{cases} 1 & \text{if } F_2 \\ 0 & \text{if not} \end{cases}$ $x_3 = \begin{cases} 1 & \text{if } F_3 \\ 0 & \text{if not} \end{cases}$ Base Level $= F_1$

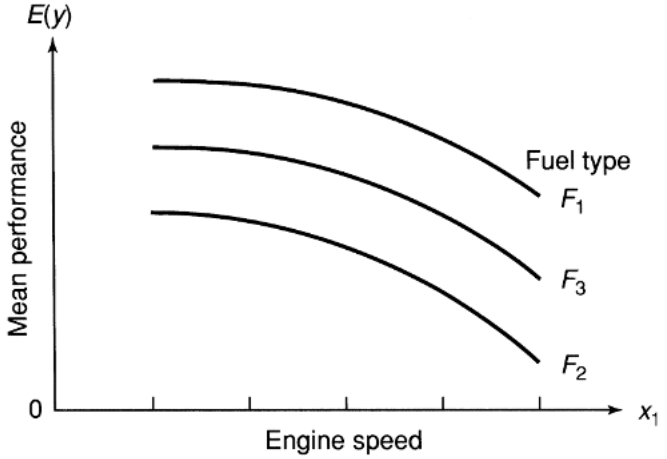


Figure 5.10: Model for $E(y)$ as a function of fuel type and engine speed (no interaction)

This non interactive second-stage model has drawbacks similar to those of the simple first-stage model. Explain.

The final stage of the model-building process — adding interaction terms to allow the three response curves to differ in shape:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1^1 x_2 + \beta_8 x_1^2 x_3$$

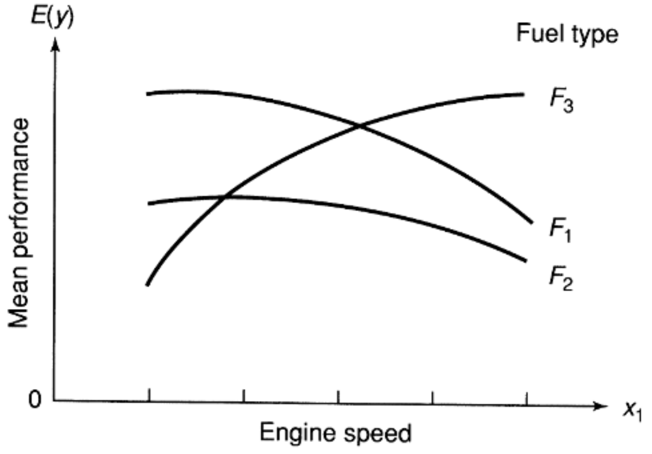


Figure 5.11: Graph of $E(y)$ as a function of fuel type and engine speed (interaction)

Question: What is the order of this model?

Example 5.5.1 *Refer to Example 5.14 from the text book.*

A marine biologist wished to investigate the effects of three factors on the level of the contaminant DDT found in fish inhabiting a polluted lake.

The variables were:

y = Level of the contaminant DDT

x_1 = Species of fish (two levels, S_1, S_2)

x_2 = Location of capture (two levels, L_1, L_2)

x_3 = Fish length (centimeters)

a) Write a model relating y to the quantitative factor(s).

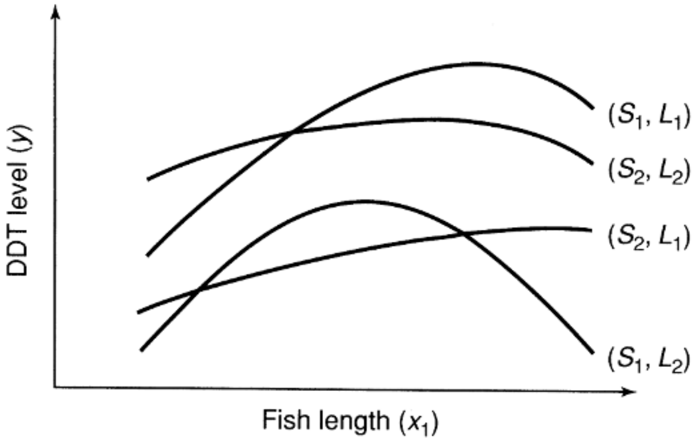


Figure 5.12: A graphical portrayal of three factors - two qualitative and one quantitative - on DDT level

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

This is the model we would use if we were certain that the DDT curves were identical for all species – location combinations (S_i , L_j). (Figure 5.13 a)

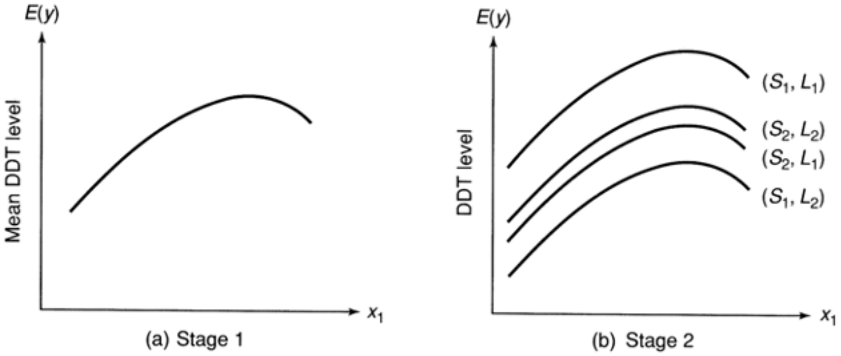


Figure 5.13: DDT curves for stages 1 and 2

b) Add the terms, both main effect and interaction, for the qualitative factors.

$$x_2 = \begin{cases} 1 & \text{if species} \quad S_2 \\ 0 & \text{if not} \end{cases} \qquad x_3 = \begin{cases} 1 & \text{if location} \quad L_2 \\ 0 & \text{if not} \end{cases}$$

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_2 x_3$$

This model implies that the DDT curves are identically shaped for each of the (Si,Lj) combinations but that they possess different y-intercepts, as shown in Figure 5.13 b.

c) Add terms to allow for interaction between the quantitative and qualitative factors

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_2 x_3 + \beta_6 x_1 x_2 + \beta_7 x_1 x_3 + \beta_8 x_1 x_2 x_3 + \beta_9 x_1^2 x_2 + \beta_{10} x_1^2 x_3 + \beta_{11} x_1^2 x_2 x_3$$

d) Use the model in part (c) to find the equation relating $E(y)$ to x_1 for species S_1 and location L_2.

Acknowledgement

The core content of the slides are from the textbook of this course;

A Second Course in Statistics: Regression Analysis (7th Edition)

by

Mendenhall, William and Sincich, Terry; Pearson Education.