CHAPTER 8

Indicator Variables

The General Concept of Indicator Variables

- **Qualitative variables** – also known as categorical variables. Qualitative variables do not have a scale of measurement
- **Indicator variables** – a variable that assigns levels to the qualitative variable (also known as **dummy** variables)

The General Concept of Indicator Variables

Example

- Relate the effective life of a cutting tool (y) used on a lathe to the lathe speed in revolutions per minute (x_1) and type of cutting tool used.
- Tool type is qualitative and can be represented as:

$$x_2 = \begin{cases} 0 & ToolA \\ 1 & ToolB \end{cases}$$

- If a first-order model is appropriate:

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

The General Concept of Indicator Variables

Example

- For Tool type A this model becomes: $y = \beta_0 + \beta_1 x_1 + \varepsilon$

- For Tool type B this model becomes:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 + \varepsilon$$

$$y = (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon$$

- Changing from A to B induces a *change* in the intercept (slope is unchanged and identical)
- It is assumed that the variance is equal for all levels of the qualitative variable

The General Concept of Indicator Variables

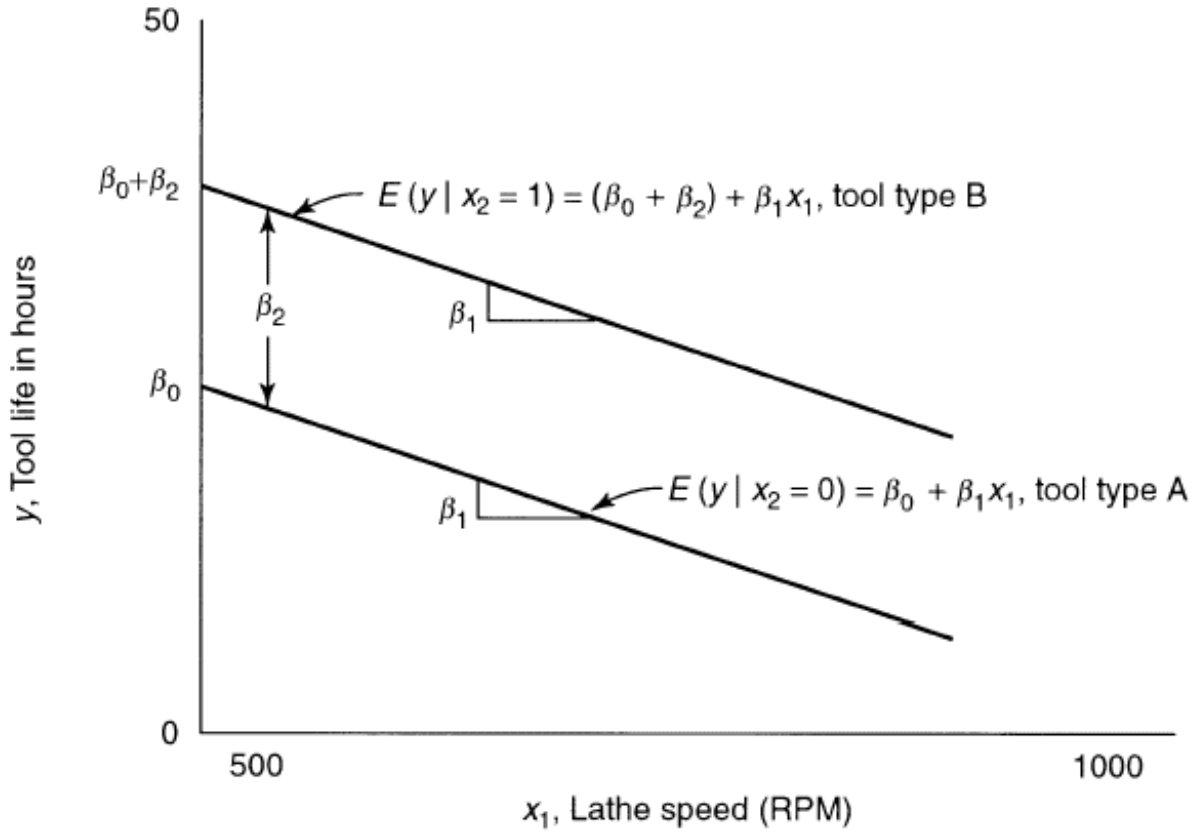

Figure 8.1 Response functions for the tool life example.

The General Concept of Indicator Variables

- For qualitative variables with ***k*** levels, we would need ***k*-1** indicator variables
- Suppose there were three tool types, A, B, and C
- Then two indicator variables (called x_2 and x_3) will be needed:

x_2	x_3	
0	0	if the observation is from tool type A
1	0	if the observation is from tool type B
0	1	if the observation is from tool type C

the regression model is

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

Example 8.1 Tool Life Data

TABLE 8.1 Data, Fitted Values, and Residuals for Example 8.1

i	y_i(Hours)	x_{i1} (rpm)	Tool Type	$\hat{y}_i$	e_i
1	18.73	610	A	20.7552	−2.0252
2	14.52	950	A	11.7087	2.8113
3	17.43	720	A	17.8284	−0.3984
4	14.54	840	A	14.6355	−0.0955
5	13.44	980	A	10.9105	2.5295
6	24.39	530	A	22.8838	1.5062
7	13.34	680	A	18.8927	−5.5527
8	22.71	540	A	22.6177	0.0923
9	12.68	890	A	13.3052	−0.6252
10	19.32	730	A	17.5623	1.7577
11	30.16	670	B	34.1630	−4.0030
12	27.09	770	B	31.5023	−4.4123
13	25.40	880	B	28.5755	−3.1755
14	26.05	1000	B	25.3826	0.6674
15	33.49	760	B	31.7684	1.7216
16	35.62	590	B	36.2916	−0.6716
17	26.07	910	B	27.7773	−1.7073
18	36.78	650	B	34.6952	2.0848
19	34.95	810	B	30.4380	4.5120
20	43.67	500	B	38.6862	4.9838

R code

- rm(list=ls())
- Life <- read.csv("data-ex-8-1.csv",h=T)
- pairs(Life,pch=20)
- # creat a dummy variable
- Life$TypeB=1*(Life$ToolType=="B")
- Life$TypeB
- # creat color vector for plotting
- col_vec=ifelse(Life$TypeB==1,"blue","red")
- pch_vec=ifelse(Life$TypeB==1,15,18)
- plot(Life$rpm,Life$Hour,col=col_vec,pch=pch_vec)
- # fit regression
- model1 <- lm(Life$Hour ~ Life$rpm+Life$TypeB)
- summary(model1)

Example 8.1 Tool Life Data

- The model to be fit is

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

where $x_2 = 0$ indicates Tool type A, if $x_2 = 1$ then Tool type B is used.

- The least squares fit is

$$\hat{y} = 36.986 - 0.027 x_1 + 15.004 x_2$$

Example 8.1 Tool Life Data

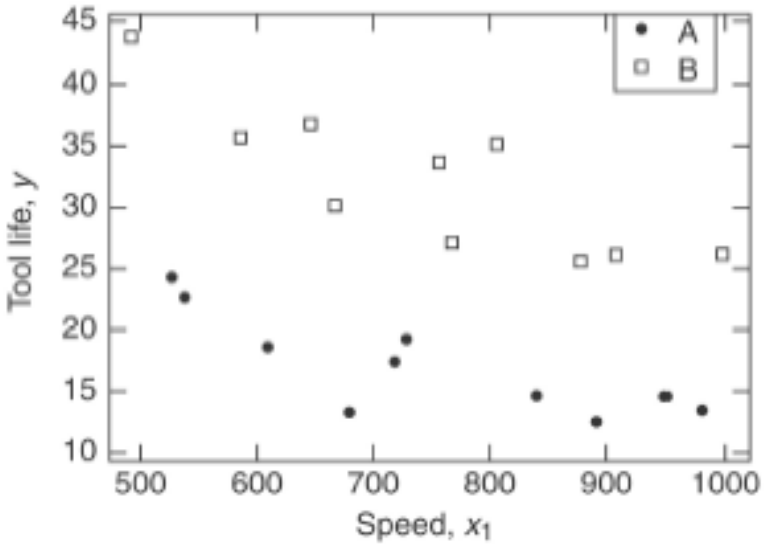

Figure 8.2 Plot of tool life y versus lathe speed x_1 for tool types A and B.

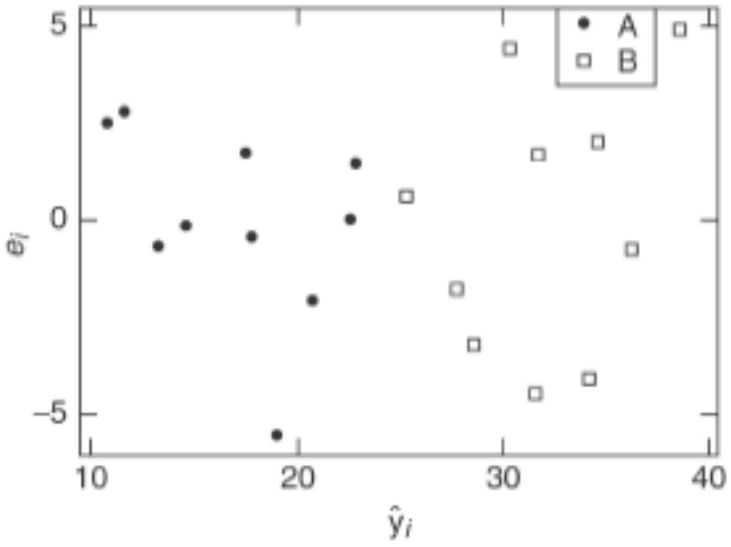

Figure 8.3 Plot of residuals e_i versus fitted values $\hat{y}_i$, Example 8.1.

Example 8.1 Tool Life Data

TABLE 8.2 Summary Statistics for the Regression Model in Example 8.1

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P Value
Regression	1418.034	2	709.017	76.75	3.12×10^{-9}
Residual	157.055	17	9.239		
Total	1575.089	19			

Coefficient	Estimate	Standard Error	t_0	P Value
β_0	36.986			
β_1	−0.027	0.005	−5.887	8.97×10^{-6}
β_2	15.004	1.360	11.035	1.79×10^{-9}
	$R^2 = 0.9003$			

Example 8.1 Tool Life Data

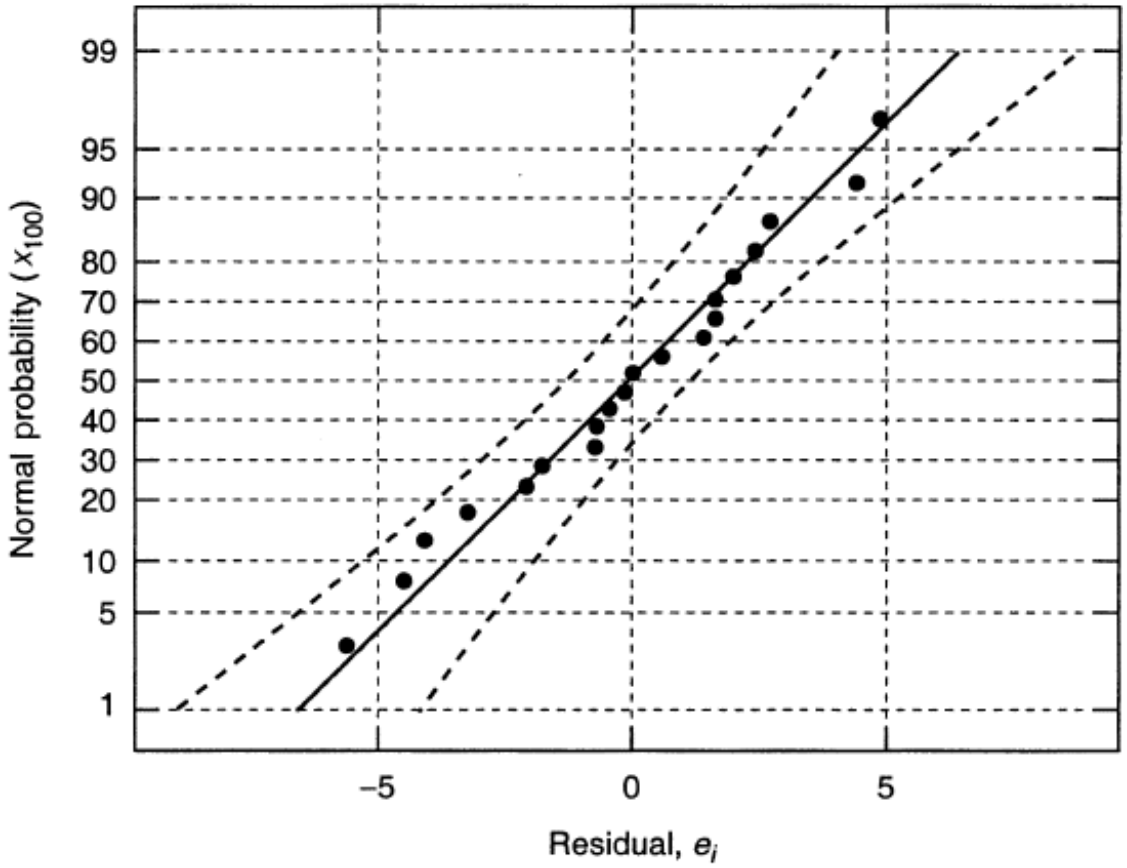

Figure 8.4 Normal probability plot of residuals, Example 8.1.

The General Concept of Indicator Variables

- Two separate models could have been fit to the data
- The single-model approach is preferred because the analyst has only one final equation
- Since the slope is assumed to be the same, it makes sense to use the data from both tool types to produce a single estimate of this common parameter

The General Concept of Indicator Variables

- If the slopes are expected differ, an interaction term can included in the model
- Considering the tool life example, the model to account for the change in slope is:

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

8.4

The General Concept of Indicator Variables

- If tool type A is used: $y = \beta_0 + \beta_1 x_1 + \varepsilon$
- If tool type B is used: $y = \beta_0 + \beta_1 x_1 + \beta_2 (1) + \beta_3 x_1 (1) + \varepsilon$

$$= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \varepsilon$$

- β_2 – change in the intercept caused by changing from type A to type B
- β_3 – change in the slope caused by changing from type A to type B

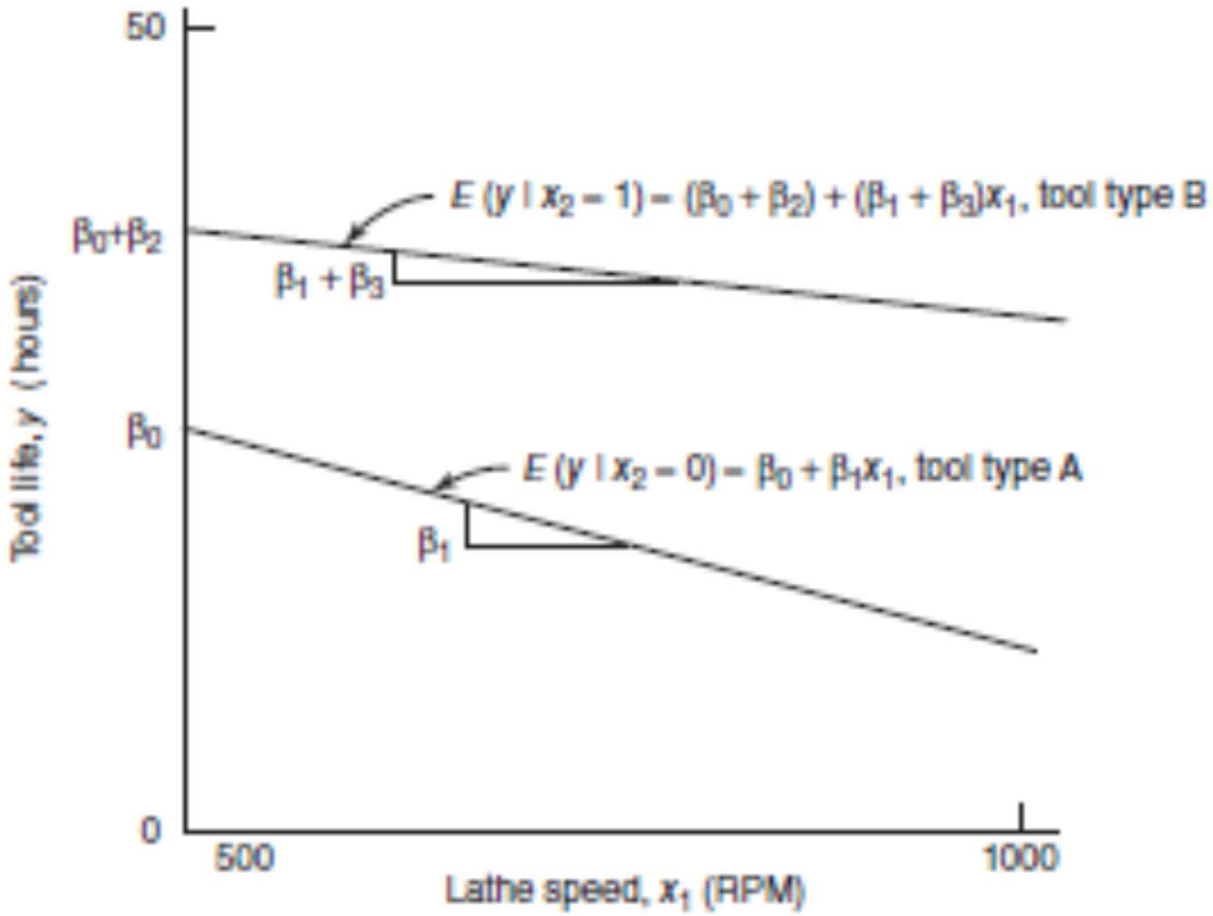

Figure 8.5 Response functions for Eq. (8.4).

R code

- # fit regression
- model2 <- lm(Life$Hour ~ Life$rpm+Life$TypeB+Life$rpm:Life$TypeB)
- summary(model2)

The General Concept of Indicator Variables

- To test to determine if these two equations are the same, use the extra sum of squares method and conduct a test of hypothesis:

$$H_0: \beta_2 = \beta_3 = 0 \text{ vs. } H_1: \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0$$

- The test statistic would be:

$$F_0 = \frac{SS_R(\beta_2, \beta_3 \mid \beta_1, \beta_0)/2}{MS_{\text{Res}}}$$

Example 8.2 The Tool Life Data

We will fit the regression model

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

The fitted regression model is

$$\hat{y} = 32.775 - 0.021 x_1 + 23.971 x_2 - 0.012 x_1 x_2$$

Example 8.2 The Tool Life Data

TABLE 8.3 Summary Analysis for the Tool Life Regression Model in Example 8.2

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P Value
Regression	1434.112	3	478.037	54.25	1.32×10^{-9}
Error	140.976	16	8.811		
Total	1575.088	19			

Coefficient	Estimate	Standard Error	t_0	Sum of Squares
β_0	32.775			
β_1	−0.021	0.0061	−3.45	$SS_R(\beta_1 \mid \beta_0) = 293.005$
β_2	23.971	6.7690	3.54	$SS_R(\beta_2 \mid \beta_1, \beta_0) = 1125.029$
β_3	−0.012	0.0088	−1.35	$SS_R(\beta_3 \mid \beta_2, \beta_1, \beta_0) = 16.078$
		$R^2 = 0.9105$		

Example 8.3 More Than Two Levels

- An electric utility is investigating the effect of the size of a single-family house and the type of air conditioning used in the house on the total electricity consumption during warm weather months

Type of Air Conditioning	x_2	x_3	x_4
No air conditioning	0	0	0
Window units	1	0	0
Heat pump	0	1	0
Central air conditioning	0	0	1

Example 8.3 More Than Two Levels

The regression model is

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

If the house has no air conditioning, Eq. (8.7) becomes

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

If the house has window units, then

$$y = (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon$$

If the house has a heat pump, the regression model is

$$y = (\beta_0 + \beta_3) + \beta_1 x_1 + \varepsilon$$

while if the house has central air conditioning, then

$$y = (\beta_0 + \beta_4) + \beta_1 x_1 + \varepsilon$$

Example 8.3 More Than Two Levels

- Appears unrealistic to assume that the slope of the regression function relating electricity consumption to the size of the house does not depend on the type of air conditioning system.
- Would expect the mean electricity consumption to increase with the size of the house, but the rate of increase would be different depending on type

Example 8.3 More Than Two Levels

- There should be an **interaction** between the size of the house and the type of air conditioning system

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \varepsilon$$

Example 8.3 More Than Two Levels

- The four regression models corresponding to the four types of air conditioning systems are as follows:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon_1 \quad \text{(no air conditioning)}$$

$$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_5) x_1 + \varepsilon \quad \text{(window units)}$$

$$y = (\beta_0 + \beta_3) + (\beta_1 + \beta_6) x_1 + \varepsilon \quad \text{(heat pump)}$$

$$y = (\beta_0 + \beta_4) + (\beta_1 + \beta_7) x_1 + \varepsilon \quad \text{(central air conditioning)}$$

Example 8.4 More than Two Indicator Variables

- Suppose that in Example 8.1 a second qualitative factor, the type of cutting oil used, must be considered.
- Assuming that this factor has two levels, we may define a second indicator variable, x_3, as follows:

$$x_3 = \begin{cases} 0 & \text{if low-viscosity oil used} \\ 1 & \text{if medium-viscosity oil used} \end{cases}$$

Example 8.4 More than Two Indicator Variables

A regression model relating tool life (y) to cutting speed (x_1), tool type (x_2), and type of cutting oil (x_3) is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \qquad (8.9)$$

Clearly the slope β_1 of the regression model relating tool life to cutting speed does not depend on either the type of tool or the type of cutting oil. The intercept of the regression line depends on these factors in an additive fashion.

Example 8.4 More than Two Indicator Variables

Various types of interaction effects may be added to the model. For example, suppose that we consider interactions between cutting speed and the two qualitative factors, so that model (8.9) becomes

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \varepsilon \qquad (8.10)$$

This implies the following situation:

Tool Type	Cutting Oil	Regression Model
A	Low viscosity	$y = \beta_0 + \beta_1 x_1 + \varepsilon$
B	Low viscosity	$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) x_1 + \varepsilon$
A	Medium viscosity	$y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) x_1 + \varepsilon$
B	Medium viscosity	$y = (\beta_0 + \beta_2 + \beta_3) + (\beta_1 + \beta_4 + \beta_5) x_1 + \varepsilon$

Example 8.4 More than Two Indicator Variables

Suppose that we add a cross-product term involving the two indicator variables x_2 and x_3 to the model, resulting in

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \varepsilon \quad (8.11)$$

We then have the following:

Tool Type	Cutting Oil	Regression Model
A	Low viscosity	$y = \beta_0 + \beta_1 x_1 + \varepsilon$
B	Low viscosity	$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1 + \varepsilon$
A	Medium viscosity	$y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1 + \varepsilon$
B	Medium viscosity	$y = (\beta_0 + \beta_2 + \beta_3 + \beta_6) + (\beta_1 + \beta_4 + \beta_5)x_1 + \varepsilon$

Some Comments on Indicator Variables

- Avoid using a specific metric for the levels of the qualitative variable (avoid allocating codes)
- Can substitute indicator variables for quantitative regressors
  - Useful if accurate data cannot be readily attained
  - Group the data into classes or intervals and assign indicator variables

Using SAS and R for Indicator Variables

- R

```
tool <- read.delim("e:\\data-ex-8-1 (Tool Life).txt",h=T)
 tool$tt <- ifelse ((tool$Tool == "A"), 0, 1)
 tool$x [tool$OBS=="7"]<- 680
 summary(temp <- lm(y ~ x+Tool, data=tool))
 influence.measures(temp)
 yhat <- temp$fit
 t <- rstudent(temp)
 par(mfrow=c(3,2))
 qqnorm(t)
 qqline(t)
 hist(t)
 plot(yhat,t)
 plot(tool$x,t)
 plot(tool$Tool,t)
 vif(temp)
```