Chapter 4

Indicator Variables as Predictors

The variables employed in regression analysis are usually quantitative variables, that is, the variables have a well-defined scale of measurement. Occasionally it is necessary to use qualitative or categorical variables as predictors. In general, a qualitative variable has no natural scale of measurement. We must assign a set of levels to a qualitative variable to account for the effect that the variable may have on the response. This is done through the use of indicator variables (also called dummy variables). In general, a qualitative variable with a levels is represented by $a-1$ indicator variables, each taking on the values 0 and 1.

Examples

- Gender (male or female)

$$X = \begin{cases} 0 & \text{if the observation is male} \\ 1 & \text{if the observation is female.} \end{cases}$$

- Shifts (day, evening, or night)

$$X_1 = \begin{cases} 1 & \text{if the observation is from day shift} \\ 0 & \text{otherwise,} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if the observation is from evening shift} \\ 0 & \text{otherwise.} \end{cases}$$

- Classification (freshmen, sophomore, junior and senior)

$$X_1 = \begin{cases} 1 & \text{if the observation is from sophomore} \\ 0 & \text{otherwise,} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if the observation is from junior} \\ 0 & \text{otherwise.} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if the observation is from senior} \\ 0 & \text{otherwise.} \end{cases}$$

The choice of 0 and 1 to identify the levels of a qualitative variable is arbitrary.

4.1 Regression with Indicator Variables

Let us begin by this section with a example. Suppose that a mechanical engineer wishes to relate the effective life of a cutting tool (Y) used on a lathe to the lathe speed in revolutions per minute (X_1) and the type of cutting tool used. The predictor, tool type, is qualitative variable and has two levels (e.g., tool types A and B). The indicator variable that is used to identity the classes of the predictor tool type is

$$X_2 = \begin{cases} 0 & \text{if the observation is from tool type A} \\ 1 & \text{if the observation is from tool type B.} \end{cases}$$

Assuming that a first-order model is appropriate, we have

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

When $X_2 = 0$, the regression model becomes

$$Y = \beta_0 + \beta_1 X_1 + \epsilon.$$

When $X_2 = 1$, the regression model becomes

$$Y = (\beta_0 + \beta_2) + \beta_1 X_1 + \epsilon.$$

That is, for tool type A the relationship between tool life and lathe speed is a straight line with intercept β_0 and slope β_1. The relationship between tool life and lathe speed for tool type B is also a straight line with slope β_1 but intercept $\beta_0 + \beta_2$. Thus, the parameter β_2 is a measure of the difference in mean tool life resulting from changing from tool type A to tool type B.

Example

See the tool life data in Ex 8.1 in textbook.

Twenty observations on tool life and lathe speed are shown in Figure 4.1.

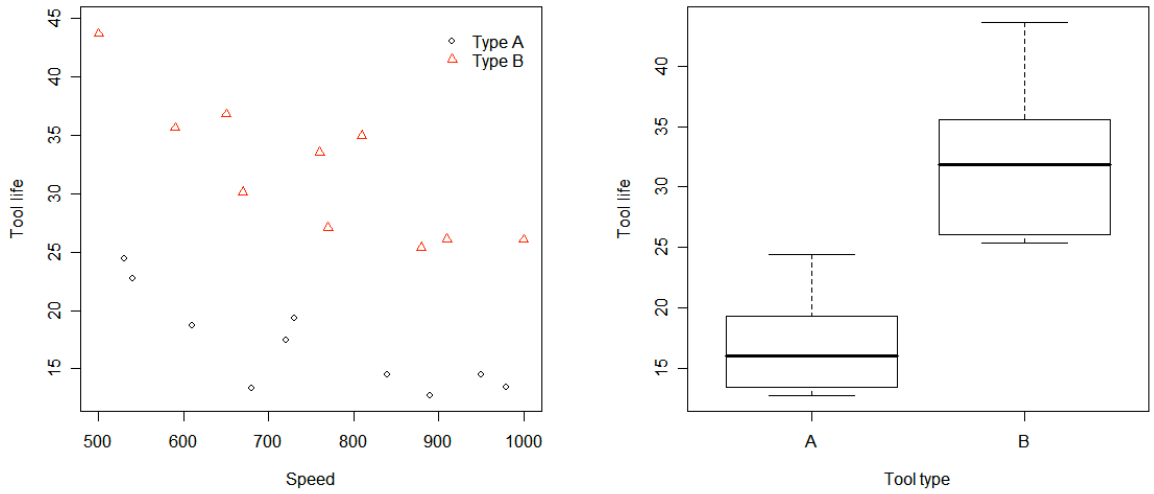

Figure 4.1: Scatter plot of tool life against lathe speed and side-by-side boxplot of tool life for tool type.

SAS Program

```
option ls=90 ps=75;
title 'Tool Life data';
data Tool;
set Toollife;
if Tooltype = 'A' then Type=0;
if Tooltype = 'B' then Type=1;
SpeedType = Speed * Type;  /* Define interaction term */
run;
/* Boxplot */
proc sort data=Tool out=sortTool; by Type; run;
proc boxplot data=sortTool; plot Toollife*Type; run;
/* Different intercept but the same slope */
proc reg data=Tool;
model Toollife = Speed Type;
run; quit;
```

Output

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1418.03362	709.01681	76.75	<.0001
Error	17	157.05456	9.23850		
Corrected Total	19	1575.08818			

Root MSE	3.03949	R-Square	0.9003
Dependent Mean	24.51900	Adj R-Sq	0.8886
Coeff Var	12.39647		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	36.98560	3.51038	10.54	<.0001
Speed	1	-0.02661	0.00452	-5.89	<.0001
Type	1	15.00425	1.35967	11.04	<.0001

Now suppose that we expect the regression lines relating tool life to lathe speed to differ in both intercept and slope. For this, the model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

Then, when $X_2 = 0$, this model becomes

$$Y = \beta_0 + \beta_1 X_1 + \epsilon.$$

When $X_2 = 1$, the regression model becomes

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + \epsilon.$$

That is, for tool type A the relationship between tool life and lathe speed is a straight line with intercept β_0 and slope β_1. The relationship between tool life and lathe speed for tool type B is also a straight line with intercept $\beta_0 + \beta_2$ and slope $\beta_1 + \beta_3$. Therefore, the parameter β_2 reflects the change in the intercept associated with changing from tool type A to tool type B and β_3 indicates the change in the slope associated with changing from tool type A to tool type B.

Fitting the regression model using indicator variables is equivalent to fitting separate regression equations. An advantage to the use of indicator variables is that tests of hypotheses can be performed directly using the partial F test.

- To test whether or not the two regression models are identical, we would test

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

- To test that the two regression models have a common slope but possibly different intercepts, we would test

$$H_0 : \beta_3 = 0 \quad \text{versus} \quad H_1 : \beta_3 \neq 0.$$

Example

Revisit the tool life data

SAS Program

```
/* Different intercept and slope model */
proc reg data=Tool;
model Toollife = Speed Type SpeedType;
run; quit;
```

Output

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1434.11235	478.03745	54.25	<.0001
Error	16	140.97583	8.81099		
Corrected Total	19	1575.08818			

Root MSE	2.96833	R-Square	0.9105
Dependent Mean	24.51900	Adj R-Sq	0.8937
Coeff Var	12.10625		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|

Intercept	1	32.77476	4.63347	7.07	<.0001
Speed	1	-0.02097	0.00607	-3.45	0.0033
Type	1	23.97059	6.76897	3.54	0.0027
SpeedType	1	-0.01194	0.00884	-1.35	0.1955

4.2 Regression Approach to Analysis of Variance

The analysis of variance is a technique frequently used to analyze data from planned or designed experiments. Although special computing procedures are generally used for analysis of variance (AOV), any AOV problem can also be treated as a linear regression problem. Ordinarily we do not recommend that regression methods be used for AOV because the specialized computing techniques are usually quite efficient. However, there are some AOV situations, particularly those involving unbalance designs and analysis of covariance (ANCOVA), where the regression approach is helpful. Essentially, any AOV problem can be treated as a regression problem in which all of the predictors are indicator variables.

In this section, we illustrate the regression alternative to the one-way AOV model.

4.2.1 Analysis of Variance

The model for the one-way AOV is

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, 2, \ldots, k, \ j = 1, 2, \ldots, r_i,$$

where Y_{ij} is the jth observation for the ith treatment or factor level, μ is a parameter common to all k treatments, τ_i is a parameter that represents the effect of the ith treatment, and ϵ_{ij} is an random error with zero mean and variance σ^2. The random errors are assumed to be normally distributed, i.e., $\epsilon_{ij} \sim i.i.d.N(0, \sigma^2)$. It is customary to define the treatment effects as

$$\tau_1 + \tau_2 + \cdots + \tau_k = 0.$$

Furthermore, the mean of the ith treatments is $\mu_i = \mu + \tau_i$, $i = 1, 2, \ldots, k$. In the fixed effects case, the AOV is used to test the hypothesis that all k population means are equal, or equivalently,

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_k = 0$$

against

$$H_1 : \tau_i \neq 0 \text{ for at least one } i.$$

Source of variation	Degree of freedom	Sum of squares	Mean squares	F_0
Treatments (or Between)	$k-1$	$SS_{TRT} = \sum_{i=1}^{k} r_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$MS_{TRT} = SS_{TRT}/(k-1)$	$\frac{MS_{TRT}}{MS_{Res}}$
Error (or Within)	$n-k$	$SS_{Res} = \sum_{i=1}^{k}\sum_{j=1}^{r_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$	$MS_{Res} = SS_{Res}/(n-k)$	
Total	$n-1$	$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{r_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$		

Table 4.1: One-Way Analysis of Variance

Here $n = \sum_{i=1}^{k} r_i$, $\bar{Y}_{i\cdot} = \frac{1}{r_i}\sum_{j=1}^{r_i} Y_{ij}$ and $\bar{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{r_i} Y_{ij}$.

The null is rejected at level α if $F_0 > F_{\alpha,k-1,n-k}$ or $P - value < \alpha$.

4.2.2 Regression Approach

To illustrate the connection between the AOV fixed effects analysis of variance and regression, suppose that we have $k = 3$ treatments, so that the AOV model becomes

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, 2, 3, \ j = 1, 2, \ldots, r_i.$$

These three treatments may be viewed as three levels of a qualitative variable, and they can be handled using indicator variables. Specifically a qualitative variable with three levels would require two indicator variables defined as follows:

$$X_1 = \begin{cases} 1 & \text{if the observation is from treatment 1} \\ 0 & \text{otherwise,} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if the observation is from treatment 2} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the regression model becomes

$$Y_{ij} = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \epsilon_{ij}, \quad i = 1, 2, 3, \ j = 1, \ldots, r_i,$$

where X_{1j} is the value of the indicator variable X_1 for observation j in treatment i and X_{2j} is the value for observation j in treatment i.

The relationship between the parameters β_0, β_1, β_2 in the regression model and the parameters μ and $\tau_i, i = 1, 2, 3$ in the AOV model is easily determined. The regression model for an observation from treatment 1 becomes

$$Y_{1j} = \beta_0 + \beta_1(1) + \beta_2(0) + \epsilon_{1j} = \beta_0 + \beta_1 + \epsilon_{1j}.$$

Since in the AOV model an observation from treatment 1 is represented by

$$Y_{1j} = \mu + \tau_i + \epsilon_{1j} = \mu_1 + \epsilon_{1j},$$

this implies that

$$\beta_0 + \beta_1 = \mu_1.$$

Similarly, the regression model for an observation from treatment 2 is $Y_{2j} = \beta_0 + \beta_2 + \epsilon_{2j}$ and the AOV model is $Y_{2j} = \mu + \tau_2 + \epsilon_{2j} = \mu_2 + \epsilon_{2j}$, and so

$$\beta_0 + \beta_2 = \mu_2.$$

Finally, considering the regression model and AOV model for an observation from treatment 3 leads

$$\beta_0 = \mu_3.$$

Thus,

$$\beta_0 = \mu_3, \quad \beta_1 = \mu_1 - \mu_3, \quad \beta_2 = \mu_2 - \mu_3.$$

That is, in the regression model formulation of the one-way AOV, the regression coefficients describe comparisons of the first two treatment means μ_1 and μ_2 with the third treatment mean μ_3.

In general, if there are k treatments, the regression model for the one-way AOV will require $k-1$ indicator variables (or dummy variables). Then, the regression model for the one-way AOV is

$$Y_{ij} = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_{k-1} X_{k-1,j} + \epsilon_{ij}, \quad i = 1, \ldots, k, \ j = 1, \ldots, r,$$

where

$$X_1 = \begin{cases} 1 & \text{if the observation is from treatment 1} \\ 0 & \text{otherwise,} \end{cases}$$

$$\vdots$$

$$X_{k-1} = \begin{cases} 1 & \text{if the observation is from treatment } k-1 \\ 0 & \text{otherwise.} \end{cases}$$

The relationship between the parameters in the regression and AOV models is

$$\begin{aligned} \beta_0 &= \mu_k \\ \beta_i &= \mu_i - \mu_k, \quad i = 1, 2, \ldots, k-1. \end{aligned}$$

Thus, β_0 always estimates the mean of the kth treatment and β_i estimates the differences in means between treatment i and treatment k.

Example

The shelf life of stored meats is the time a prepackaged cut remains salable, safe, and nutritious. Standard packaging in ambient air atmosphere has a shelf life of about 48 hours after which the meat quality begins to deteriorate from microbial contamination, color degradation, and shrinkage. Vacuum packaging is effective in suppressing microbial growth; however, other quality losses remain a problem. Recent studies suggested controlled gas atmospheres are possible alternatives to existing packaging. Two atmospheres which promise to combine the capability for suppressing microbial development while maintaining other meat qualities were pure carbon dioxide (CO2), and (2) mixture of carbon monoxide (CO), oxygen (O2), and nitrogen (N). Three beef steaks of approximately the same size (75g) were randomly assigned to each of the packaging conditions. Each steak was packaged separately in its assigned conditions. The number of psychrotrophic bacteria on the meat was measured after nine days of storage at 4°C in a standard meat storage facility. Psychrotrophic bacteria are found on the surface of the meat and are associated with spoilage of the meat product.

	Psychrotrophic Bacteria		
Packaging Condition	Log(count/cm^2)		
Commercial plastic wrap	7.66	6.98	7.80
Vacuum packaged	5.26	5.44	5.80
1% CO, 40% O2, 59% N	7.41	7.33	7.04
100% CO2	3.51	2.91	3.66

Based on this new information, the researcher wanted to see that some form of controlled gas atmosphere would provide a more effective packaging environment for meat storage. Test the hypothesis of no difference among the mean bacterial counts (measured in log scale) of the packaging conditions.

SAS Output

Obs	packaging	bacteria	x1	x2	x3
1	Commerci	7.66	1	0	0
2	Commerci	6.98	1	0	0
3	Commerci	7.80	1	0	0
4	Vacuum	5.26	0	1	0
5	Vacuum	5.44	0	1	0
6	Vacuum	5.80	0	1	0
7	Mixed	7.41	0	0	1
8	Mixed	7.33	0	0	1
9	Mixed	7.04	0	0	1
10	PureCO2	3.51	0	0	0
11	PureCO2	2.91	0	0	0
12	PureCO2	3.66	0	0	0

The REG Procedure
Model: MODEL1
Dependent Variable: bacteria

Number of Observations Read	12
Number of Observations Used	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	32.87280	10.95760	94.58	<.0001
Error	8	0.92680	0.11585		
Corrected Total	11	33.79960			

Root MSE	0.34037	R-Square	0.9726
Dependent Mean	5.90000	Adj R-Sq	0.9623
Coeff Var	5.76894		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.36000	0.19651	17.10	<.0001
x1	1	4.12000	0.27791	14.83	<.0001
x2	1	2.14000	0.27791	7.70	<.0001
x3	1	3.90000	0.27791	14.03	<.0001

The ANOVA Procedure

Class Level Information

Class	Levels	Values
packaging	4	Commerci Mixed PureCO2 Vacuum

Number of Observations Read	12
Number of Observations Used	12

Dependent Variable: bacteria

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	32.87280000	10.95760000	94.58	<.0001
Error	8	0.92680000	0.11585000		
Corrected Total	11	33.79960000			

R-Square	Coeff Var	Root MSE	bacteria Mean
0.972580	5.768940	0.340367	5.900000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
packaging	3	32.87280000	10.95760000	94.58	<.0001

Level of packaging	N	bacteria Mean	bacteria Std Dev
Commerci	3	7.48000000	0.43863424
Mixed	3	7.26000000	0.19467922
PureCO2	3	3.36000000	0.39686270
Vacuum	3	5.50000000	0.27495454