

## Example 61.3: Regression with Quantitative and Qualitative Variables

At times it is desirable to have independent variables in the model that are qualitative rather than quantitative. This is easily handled in a regression framework. Regression uses qualitative variables to distinguish between populations. There are two main advantages of fitting both populations in one model. You gain the ability to test for different slopes or intercepts in the populations, and more degrees of freedom are available for the analysis.

Regression with qualitative variables is different from analysis of variance and analysis of covariance. Analysis of variance uses qualitative independent variables only. Analysis of covariance uses quantitative variables in addition to the qualitative variables in order to account for correlation in the data and reduce MSE; however, the quantitative variables are not of primary interest and merely improve the precision of the analysis.

Consider the case where  $Y_i$  is the dependent variable,  $X1_i$  is a quantitative variable,  $X2_i$  is a qualitative variable taking on values 0 or 1, and  $X1_iX2_i$  is the interaction. The variable  $X2_i$  is called a dummy, binary, or indicator variable. With values 0 or 1, it distinguishes between two populations. The model is of the form

$$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X1_i X2_i + \epsilon_i$$

for the observations  $i = 1, 2, \dots, n$ . The parameters to be estimated are  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$ . The number of dummy variables used is one less than the number of qualitative levels. This yields a nonsingular  $X'X$  matrix. See Chapter 10 of Neter, Wasserman, and Kutner (1990) for more details.

An example from Neter, Wasserman, and Kutner (1990) follows. An economist is investigating the relationship between the size of an insurance firm and the speed at which they implement new insurance innovations. He believes that the type of firm may affect this relationship and suspects that there may be some interaction between the size and type of firm. The dummy variable in the model allows the two firms to have different intercepts. The interaction term allows the firms to have different slopes as well.

In this study,  $Y_i$  is the number of months from the time the first firm implemented the innovation to the time it was implemented by the  $i$ th firm. The variable  $X1_i$  is the size of the firm, measured in total assets of the firm. The variable  $X2_i$  denotes the firm type and is 0 if the firm is a mutual fund company and 1 if the firm is a stock company. The dummy variable allows each firm type to have a different intercept and slope.

The previous model can be broken down into a model for each firm type by plugging in the values for  $X2_i$ . If  $X2_i=0$ , the model is

$$Y_i = \beta_0 + \beta_1 X1_i + \epsilon_i$$

This is the model for a mutual company. If  $X2_i=1$ , the model for a stock firm is

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X1_i + \epsilon_i$$

This model has intercept  $\beta_0 + \beta_2$  and slope  $\beta_1 + \beta_3$ .

The data\* follow. Note that the interaction term is created in the DATA step since polynomial effects such as size\*type are not allowed in the MODEL statement in the REG procedure.

```

title 'Regression With Quantitative and Qualitative Variables';
data insurance;
  input time size type @@;
  sizetype=size*type;
  datalines;
17 151 0   26  92 0   21 175 0   30  31 0   22 104 0
 0 277 0   12 210 0   19 120 0   4 290 0   16 238 0
28 164 1   15 272 1   11 295 1   38  68 1   31  85 1
21 224 1   20 166 1   13 305 1   30 124 1   14 246 1
;
run;

```

The following statements begin the analysis:

```

proc reg data=insurance;
  model time = size type sizetype;
run;

```

The ANOVA table is displayed in [Output 61.3.1](#).

**Output 61.3.1:** ANOVA Table and Parameter Estimates

Regression With Quantitative and Qualitative Variables					
The REG Procedure					
Model: MODEL1					
Dependent Variable: time					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1504.41904	501.47301	45.49	<.0001
Error	16	176.38096	11.02381		
Corrected Total	19	1680.80000			

<b>Root MSE</b>	3.32021	<b>R-Square</b>	0.8951
<b>Dependent Mean</b>	19.40000	<b>Adj R-Sq</b>	0.8754
<b>Coeff Var</b>	17.11450		

<b>Parameter Estimates</b>					
<b>Variable</b>	<b>DF</b>	<b>Parameter Estimate</b>	<b>Standard Error</b>	<b>t Value</b>	<b>Pr &gt;  t </b>
<b>Intercept</b>	<b>1</b>	33.83837	2.44065	13.86	<.0001
<b>size</b>	<b>1</b>	-0.10153	0.01305	-7.78	<.0001
<b>type</b>	<b>1</b>	8.13125	3.65405	2.23	0.0408
<b>sizetype</b>	<b>1</b>	-0.00041714	0.01833	-0.02	0.9821

The overall  $F$  statistic is significant ( $F=45.490, p<0.0001$ ). The interaction term is not significant ( $t=-0.023, p=0.9821$ ). Hence, this term should be removed and the model re-fitted, as shown in the following statements.

```
delete sizetype;
print;
run;
```

The DELETE statement removes the interaction term (sizetype) from the model. The new ANOVA table is shown in [Output 61.3.2](#).

**Output 61.3.2:** ANOVA Table and Parameter Estimates

<b>Regression With Quantitative and Qualitative Variables</b>					
The REG Procedure					
Model: MODEL1.1					
Dependent Variable: time					
<b>Analysis of Variance</b>					
<b>Source</b>	<b>DF</b>	<b>Sum of Squares</b>	<b>Mean Square</b>	<b>F Value</b>	<b>Pr &gt; F</b>
<b>Model</b>	2	1504.41333	752.20667	72.50	<.0001
<b>Error</b>	17	176.38667	10.37569		

<b>Corrected Total</b>	19	1680.80000			
<b>Root MSE</b>	3.22113	<b>R-Square</b>	0.8951		
<b>Dependent Mean</b>	19.40000	<b>Adj R-Sq</b>	0.8827		
<b>Coeff Var</b>	16.60377				
<b>Parameter Estimates</b>					
<b>Variable</b>	<b>DF</b>	<b>Parameter Estimate</b>	<b>Standard Error</b>	<b>t Value</b>	<b>Pr &gt;  t </b>
<b>Intercept</b>	<b>1</b>	33.87407	1.81386	18.68	<.0001
<b>size</b>	<b>1</b>	-0.10174	0.00889	-11.44	<.0001
<b>type</b>	<b>1</b>	8.05547	1.45911	5.52	<.0001

The overall  $F$  statistic is still significant ( $F=72.497, p<0.0001$ ). The intercept and the coefficients associated with size and type are significantly different from zero ( $t=18.675, p<0.0001$ ;  $t=-11.443, p<0.0001$ ;  $t=5.521, p<0.0001$ , respectively). Notice that the  $R^2$  did not change with the omission of the interaction term.

The fitted model is

$$\text{time} = 33.87 - 0.102 \times \text{size} + 8.055 \times \text{type}$$

The fitted model for a mutual fund company ( $X_{2i}=0$ ) is

$$\text{time} = 33.87 - 0.102 \times \text{size}$$

and the fitted model for a stock company ( $X_{2i}=1$ ) is

$$\text{time} = (33.87 + 8.055) - 0.102 \times \text{size}$$

So the two models have different intercepts but the same slope.

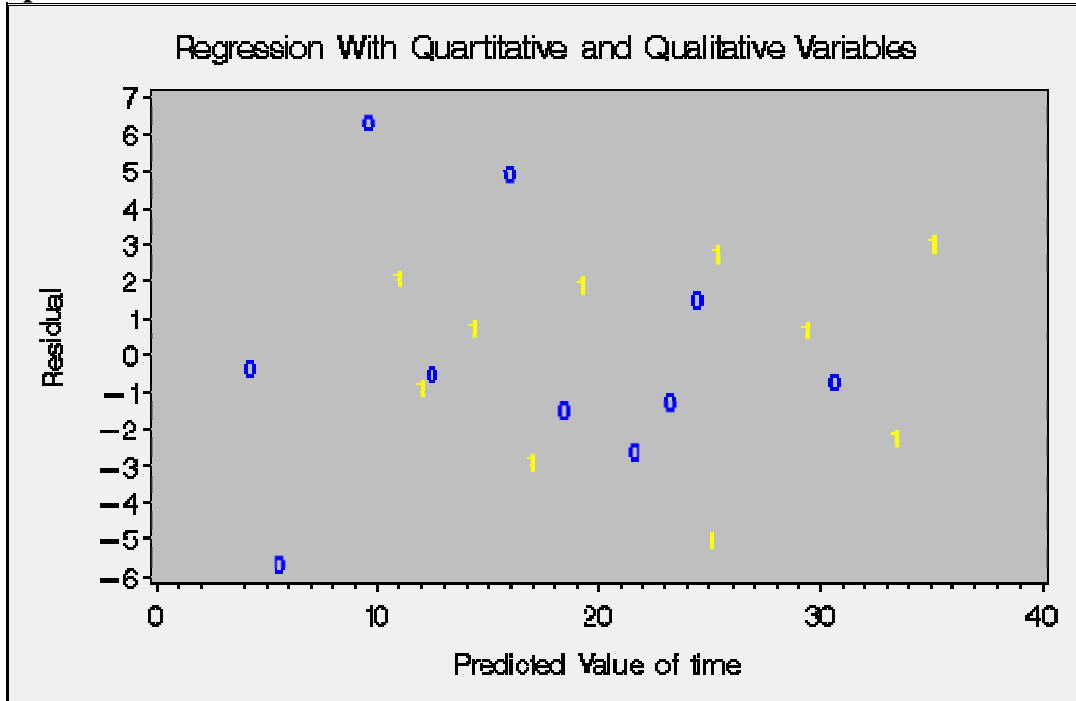
Now plot the residual versus predicted values using the firm type as the plot symbol (PLOT=TYPE); this can be useful in determining if the firm types have different residual patterns. PROC REG does not support the `plot y*x=type` syntax for high-resolution graphics, so use PROC GPLOT to create [Output 61.3.3](#). First, the OUTPUT statement saves the residuals and predicted values from the new model in the OUT= data set.

```

output out=out r=r p=p;
run;
symbol1 v='0' c=blue f=swissb;
symbol2 v='1' c=yellow f=swissb;
axis1 label=(angle=90);
proc gplot data=out;
  plot r*p=type / nolegend vaxis=axis1 cframe=ligr;
  plot p*size=type / nolegend vaxis=axis1 cframe=ligr;
run;

```

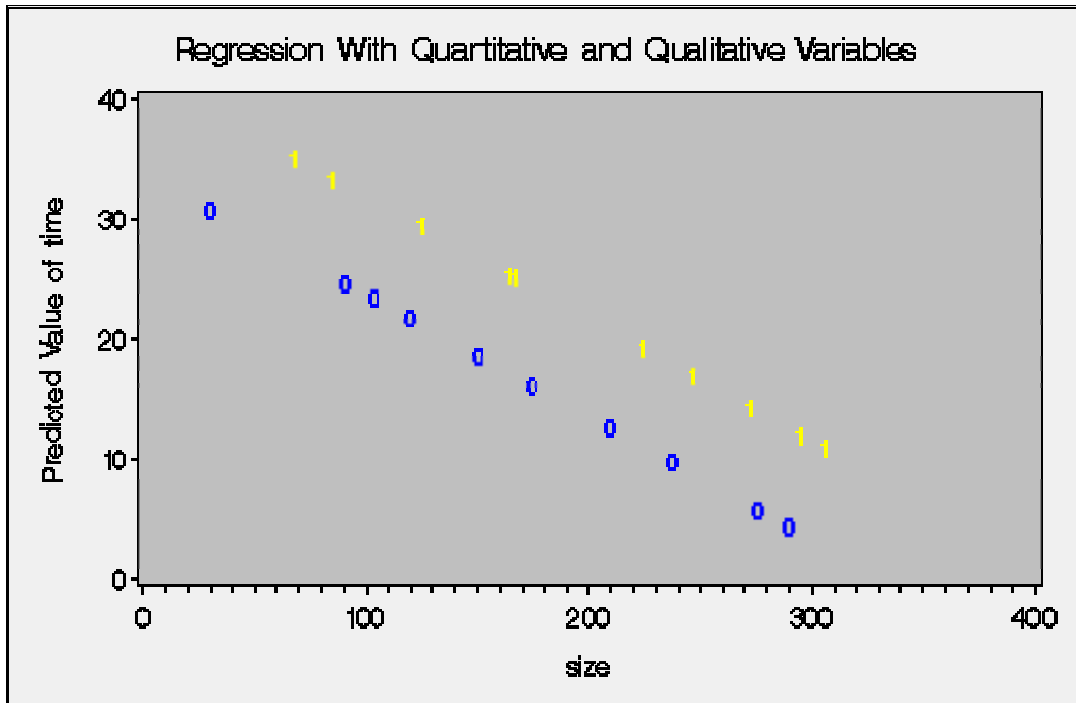
**Output 61.3.3:** Plot of Residual vs. Predicted Values



The residuals show no major trend. Neither firm type by itself shows a trend either. This indicates that the model is satisfactory.

A plot of the predicted values versus size appears in [Output 61.3.4](#), where the firm type is again used as the plotting symbol.

**Output 61.3.4:** Plot of Predicted vs. Size



The different intercepts are very evident in this plot.

[Previous](#) | [Next](#) | [Top of Page](#)

[Copyright © 2003 by SAS Institute Inc., Cary, NC, USA. All rights reserved.](#)