Using Standard Statistical Software Packages to Make Statistical Inferences About a Linear Combination of Regression Parameters

Bruce Bowerman[1], J. B. Orris[2]
Richard T. Farmer School of Business Administration, Miami University, Oxford, OH 45056[1]
College of Business Administration, Butler University, Indianapolis, IN 46208[2]

Abstract

In regression analysis, sophisticated statistical software packages such as SAS can make statistical inferences about both mean responses and about a linear combination of regression parameters. Other packages -- such as Minitab and Excel add-in packages -- make statistical inferences about mean responses but not about a linear combination of regression parameters. In this paper we show how to use a statistical software package that makes statistical inferences about mean responses to also make statistical inferences about a linear combination of regression parameters. This will be demonstrated by writing the regression model in a special way.

Keywords: Regression, MegaStat, linear combination, statistical inference

1. Introduction

Consider the multiple regression model: $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$. Standard statistical software packages such as MINITAB and Excel add-in packages such as MegaStat ®[1] compute a confidence interval for a mean response: $\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ but not for a linear combination of regression parameters: $l = l_0\beta_0 + l_1\beta_1 + \ldots + l_k\beta_k$. These packages do, however, allow you to omit the intercept, β_0. If we omit the intercept and add in a "variable intercept" we can use these standard software packages to find a confidence interval for *l*.

1.1 The Variable Intercept Model

To use a "variable intercept" we consider the model $Y = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$ where we set the variable x0 to 1 when we fit the model to obtain the usual intercept. Using this model, standard software packages find the following 100(1-α) percent confidence interval for the mean $\mu = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_k x_k$:

$$\hat{y} \pm t_{\frac{\alpha}{2}} s\sqrt{x'\left(X'X\right)^{-1} x}$$

Where: $\hat{y} = b_0 x_0 + b_1 x_1 + \ldots + b_k x_k$ and $x' = \left[x_0\ x_1 \ldots x_k\right]$. To obtain the confidence interval we specify the predictor values $x_0\ x_1 \ldots x_k$ where x0 = 1 if we assume the usual intercept. The formula for a 100(1-α) percent confidence interval for $l = l_0\beta_0 + l_1\beta_1 + \ldots + l_k\beta_k$ is:

$$\hat{l} \pm t_{\frac{\alpha}{2}} s\sqrt{l'\left(X'X\right)^{-1} l}$$

Where $\hat{l} = l_0 b_0 + l_1 b_1 + \ldots + l_k b_k$ and $l' = \left[l_0\ l_1 \ldots l_k\right]$. This formula has exactly the same form as the formula for a confidence interval for μ. Therefore, we can obtain a confidence interval for l by specifying the predictor values $x_0 = l_0, x_1 = l_1, \ldots x_k = l_k$.

2. Example

Consider Table 1, which gives values of *y*, demand for Fresh Liquid Laundry Detergent, $x_4 = x_2 - x_1$, the difference between the average industry (competitors') price and the price for Fresh, and x_3, the advertising expenditure for Fresh. To ultimately increase the demand for Fresh, Enterprise Industries' marketing department is conducting a study comparing the effectiveness of three different advertising campaigns. These campaigns are denoted as campaigns A, B, and C. Here campaign A consists entirely of television commercials, campaign B consists of a balanced mixture of television and radio commercials, and campaign C consists of a balanced mixture of television, radio, newspaper, and magazine ads. To conduct the study, Enterprise Industries has randomly selected one advertising campaign to be used in each of the 30 sales periods. Here, although logic would indicate that each of campaigns A, B. and C should be used in 10 of the 30 sales periods, Enterprise Industries has made previous commitments to the advertising media involved in the study. As a result, campaigns A, B, and C were randomly assigned to, respectively 9, 11, and 10 sales periods. Table 1 lists the campaigns used in the sales periods. To compare the effectiveness of advertising campaigns A, B, and C, we define two dummy variables. Specifically, we define D_B to equal 1 if campaign B is used in a sales period and 0 otherwise. Furthermore, we define the dummy variable D_C to equal 1 is campaign C is used in a sales period and 0 otherwise.

A possible model describing these data is then

$$y = \beta_0 x_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3$$
$$+ \beta_5 D_B + \beta_6 D_C + \beta_7 x_3 D_B + \beta_8 x_2 D_C + \varepsilon$$

where we set $x_0 = 1$ when we fit the model to obtain the usual intercept. In order to compare the advertising campaigns, consider comparing three means denoted $\mu_{[d,a,A]}$, $\mu_{[d,a,B]}$, and $\mu_{[d,a,C]}$. These means represent the mean demands for Fresh when the price difference is *d*, the advertising expenditure is *a*, and we use advertising campaigns A, B, and C, respectively. The above model implies that

$$\mu_{[d,a,A]} = \beta_0 + \beta_1 d + \beta_2 a + \beta_3 a^2 + \beta_4 da + \beta_5 (0) + \beta_6 (0) + \beta_7 a(0) + \beta_8 a(0)$$
$$\mu_{[d,a,B]} = \beta_0 + \beta_1 d + \beta_2 a + \beta_3 a^2 + \beta_4 da + \beta_5 (1) + \beta_6 (0) + \beta_7 a(1) + \beta_8 a(0)$$
$$\mu_{[d,a,C]} = \beta_0 + \beta_1 d + \beta_2 a + \beta_3 a^2 + \beta_4 da + \beta_5 (0) + \beta_6 (1) + \beta_7 a(0) + \beta_8 a(1)$$

Using these equations we find that

$$\mu_{[d,a,C]} - \mu_{[d,a,A]} = \beta_6 + \beta_8 a$$
$$\mu_{[d,a,C]} - \mu_{[d,a,B]} = \beta_6 - \beta_5 + \beta_8 a - \beta_7 a$$

We will now use the Excel add-in MegaStat to find the 95% confidence intervals for:

1. $\mu_{[d,6.2,C]} - \mu_{[d,6.2,A]} = \beta_6 + \beta_8(6.2)$
   $= 0\beta_0+0\beta_1+0\beta_2+0\beta_3+0\beta_4+0\beta_5+1\beta_6+0\beta_7+6.2\beta_8$
   Therefore we use the following predictor values:
   [0 0 0 0 0 0 1 0 6.2]

2. $\mu_{[d,6.6,C]} - \mu_{[d,6.6,A]} = \beta_6 + \beta_8(6.6)$
   $= 0\beta_0+0\beta_1+0\beta_2+0\beta_3+0\beta_4+0\beta_5+1\beta_6+0\beta_7+6.6\beta_8$
   Therefore we use the following predictor values:
   [0 0 0 0 0 0 1 0 6.6]

3. $\mu_{[d,6.2,C]} - \mu_{[d,6.2,B]} = \beta_6 - \beta_5 + \beta_8(6.2) - \beta_7(6.2)$
   $= 0\beta_0+0\beta_1+0\beta_2+0\beta_3+0\beta_4+(-1)\beta_5+1\beta_6+(-6.2)\beta_7+6.2\beta_8$
   Therefore we use the following predictor values:
   [0 0 0 0 0 -1 1-6.2 6.2]

4. $\mu_{[d,6.6,C]} - \mu_{[d,6.6,B]} = \beta_6 - \beta_5 + \beta_8(6.6) - \beta_7(6.6)$
   $= 0\beta_0+0\beta_1+0\beta_2+0\beta_3+0\beta_4+(-1)\beta_5+1\beta_6+(-6.6)\beta_7+6.6\beta_8$
   Therefore we use the following predictor values:
   [0 0 0 0 0 -1 1-6.6 6.6]

Figure 1 gives the MegaStat output of fitting the variable intercept model and the 95 percent confidence intervals.

Table 1: The Fresh Detergent Demand Data

	Price	IndPrice	AdvExp	PriceDif	Demand	AdCamp
Sales Period	x1	x2	x3	x4=x2-x1	y	
1	3.85	3.80	5.50	-0.05	7.38	B
2	3.75	4.00	6.75	0.25	8.51	B
3	3.70	4.30	7.25	0.60	9.52	B
4	3.70	3.70	5.50	0.00	7.50	A
5	3.60	3.85	7.00	0.25	9.33	C
6	3.60	3.80	6.50	0.20	8.28	A
7	3.60	3.75	6.75	0.15	8.75	C
8	3.80	3.85	5.25	0.05	7.87	C
9	3.80	3.65	5.25	-0.15	7.10	B
10	3.85	4.00	6.00	0.15	8.00	C
11	3.90	4.10	6.50	0.20	7.89	A
12	3.90	4.00	6.25	0.10	8.15	C
13	3.70	4.10	7.00	0.40	9.10	C
14	3.75	4.20	6.90	0.45	8.86	A
15	3.75	4.10	6.80	0.35	8.90	B
16	3.80	4.10	6.80	0.30	8.87	B
17	3.70	4.20	7.10	0.50	9.26	B
18	3.80	4.30	7.00	0.50	9.00	A
19	3.70	4.10	6.80	0.40	8.75	B
20	3.80	3.75	6.50	-0.05	7.95	B
21	3.80	3.75	6.25	-0.05	7.65	C
22	3.75	3.65	6.00	-0.10	7.27	A
23	3.70	3.90	6.50	0.20	8.00	A
24	3.55	3.65	7.00	0.10	8.50	A
25	3.60	4.10	6.80	0.50	8.75	A
26	3.65	4.25	6.80	0.60	9.21	B
27	3.70	3.65	6.50	-0.05	8.27	C
28	3.75	3.75	5.75	0.00	7.67	B
29	3.80	3.85	5.80	0.05	7.93	C
30	3.70	4.25	6.80	0.55	9.26	C

References

1 © 2005 by J. B. Orris. Published with selected McGraw-Hill textbooks.

Figure 1. MegaStat output of 95% Confidence Intervals for the Differences.

Regression Analysis

R^2	1.000	(Note: No intercept in the model. Interpret R^2 and R with caution.)	
Adjusted R^2	1.000	n	30
R	1.000	k	9
Std. Error	0.129	Dep. Var.	**Demand**

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Regression	2,121.1798	9	235.6866	14070.80	1.27E-37
Residual	0.3518	21	0.0168		
Total	2,121.5316	30			

Regression output

						confidence interval	
variables		*coefficients*	*std. error*	*t (df=21)*	*p-value*	*95% lower*	*95% upper*
(No Intercept)							
ones	b0	28.6873	5.1285	5.594	1.50E-05	18.0221	39.3526
PriceDif	b1	10.8253	3.2988	3.282	.0036	3.9651	17.6855
AdvExp	b2	-7.4115	1.6617	-4.460	.0002	-10.8671	-3.9558
AdvExp²	b3	0.6458	0.1346	4.798	.0001	0.3659	0.9257
PriceDif*Adv	b4	-1.4156	0.4929	-2.872	.0091	-2.4406	-0.3907
DB	b5	-0.4807	0.7309	-0.658	.5179	-2.0007	1.0393
DC	b6	-0.9351	0.8357	-1.119	.2758	-2.6731	0.8029
AdvExp*DB	b7	0.1072	0.1117	0.960	.3480	-0.1251	0.3395
AdvExp*DC	b8	0.2035	0.1288	1.580	.1291	-0.0644	0.4714

Predicted values for: Demand

ones	*PriceDif*	*AdvExp*	*AdvExp²*	*PriceDif*Adv*	*DB*	*DC*	*AdvExp*DB*	*AdvExp*DC*
0	0.0	0.0	0.0	0.0	0	1	0.0	6.2
0	0.0	0.0	0.0	0.0	0	1	0.0	6.6
0	0.0	0.0	0.0	0.0	-1	1	-6.2	6.2
0	0.0	0.0	0.0	0.0	-1	1	-6.6	6.6

		95% Confidence Intervals	
	Predicted	*lower*	*upper*
$\mu_{[d,6.2,C]} - \mu_{[d,6.2,A]}$	0.32654	0.18069	0.47240
$\mu_{[d,6.6,C]} - \mu_{[d,6.6,A]}$	0.40794	0.27668	0.53920
$\mu_{[d,6.2,C]} - \mu_{[d,6.2,B]}$	0.14245	-0.00312	0.28802
$\mu_{[d,6.6,C]} - \mu_{[d,6.6,B]}$	0.18095	0.04688	0.31503