Dummy Variables In Regression

Applied Regression and Other Multivariable Methods
Sections 14-1 - 14-12

Overview

- Have already utilized dummy variables in Topic 9
- Smoking was classified based on past history
- Converted categorical response into numeric response
- SMK = 1 if past history / SMK=0 if no past history
- Numeric response has no meaning except to define the categorical response.
- As alternative, could have used SMK = -1 or 1 without altering results (except for SMK coef and intercept)
- This topic will focus on the use of dummy variables to compare regression lines
- Will also utilize dummy variables concept in ANOVA

Creating Dummy Variables

- There is one general rule to follow

  If categorical variable has k categories and the model has an intercept, then exactly $k-1$ dummy variables are created to define them. If the model does not have an intercept, then k dummy variables should be created to define them.

- Example: Blood Type

$$Z_1 = \begin{cases} 1 \text{ if Type AB} \\ 0 \text{ otherwise} \end{cases} \qquad Z_1^\star = \begin{cases} 1 \text{ if Type AB} \\ \text{-1 if Type O} \\ 0 \text{ otherwise} \end{cases}$$

$$Z_2 = \begin{cases} 1 \text{ if Type A} \\ 0 \text{ otherwise} \end{cases} \qquad Z_2^\star = \begin{cases} 1 \text{ if Type A} \\ \text{-1 if Type O} \\ 0 \text{ otherwise} \end{cases}$$

$$Z_3 = \begin{cases} 1 \text{ if Type B} \\ 0 \text{ otherwise} \end{cases} \qquad Z_3^\star = \begin{cases} 1 \text{ if Type B} \\ \text{-1 if Type O} \\ 0 \text{ otherwise} \end{cases}$$

(Z_1, Z_2, Z_3)	Coding	$(Z_1^\star, Z_2^\star, Z_3^\star)$
$(1,0,0)$	AB	$(1,0,0)$
$(0,1,0)$	A	$(0,1,0)$
$(0,0,1)$	B	$(0,0,1)$
$(0,0,0)$	O	$(-1,-1,-1)$

Understanding Dummy Variables

- Diff coding schemes $\leftrightarrow$ diff interpretation of coefs
- Consider coding scheme 1

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3$$

AB:	$\beta_0 + \beta_1(1) + \beta_2(0) + \beta_3(0)$	$\overline{\text{AB}}$ estimates $\beta_0 + \beta_1$
A:	$\beta_0 + \beta_1(0) + \beta_2(1) + \beta_3(0)$	$\overline{\text{A}}$ estimates $\beta_0 + \beta_2$
B:	$\beta_0 + \beta_1(0) + \beta_2(0) + \beta_3(1)$	$\overline{\text{B}}$ estimates $\beta_0 + \beta_3$
O:	$\beta_0 + \beta_1(0) + \beta_2(0) + \beta_3(0)$	$\overline{\text{0}}$ estimates β_0

- Consider coding scheme 2

$$Y = \beta_0 + \beta_1 Z_1^\star + \beta_2 Z_2^\star + \beta_3 Z_3^\star$$

AB:	$\beta_0 + \beta_1(1) + \beta_2(0) + \beta_3(0)$	$\overline{\text{AB}}$ estimates $\beta_0 + \beta_1$
A:	$\beta_0 + \beta_1(0) + \beta_2(1) + \beta_3(0)$	$\overline{\text{A}}$ estimates $\beta_0 + \beta_2$
B:	$\beta_0 + \beta_1(0) + \beta_2(0) + \beta_3(1)$	$\overline{\text{B}}$ estimates $\beta_0 + \beta_3$
O:	$\beta_0 - \beta_1(1) - \beta_2(1) - \beta_3(1)$	$\overline{\text{O}}$ estimates $\beta_0 - \beta_1 - \beta_2 - \beta_3$

The intercept is the mean of the O group is coding scheme 1 while the intercept is the grand mean in coding scheme 2 (if equal number of observations of each Type).

SAS Example

Following analyses use the following dummy variable classifications for smoking history

$$\text{SMK} = \begin{cases} 1 & \text{if history} \\ 0 & \text{otherwise} \end{cases} \qquad \text{SMK1} = \begin{cases} 1 & \text{if history} \\ -1 & \text{otherwise} \end{cases}$$

A two sample t-test and the regressions SBP=SMK and SBP=SMK1 are performed

```
options nocenter;
goptions reset=global colors=(none);

data problem81;
 infile 'i:\.www\datasets502\EX0502.DAT' firstobs=2 dlm='09'x;
 input person sbp quet age smk;
 if smk = 0 then smk1 = -1;
 if smk = 1 then smk1 = 1;

proc sort; by smk;
proc ttest;
 var sbp;
 class smk;

proc reg;
 model sbp = smk;
 model sbp = smk1;

run;
quit;
```



```
The TTEST Procedure
                                   Statistics
                          Lower CL Upper CL
Variable  smk    N   Mean    Mean     Mean   Std Err
sbp       0     15 133.66   140.8   147.94    3.3312
sbp       1     17    140  147.82   155.64    3.6894
sbp     Diff (1-2)  -17.28  -7.024  3.2358    5.0235

                             T-Tests
Variable  Method          Variances      DF    t Value    Pr > |t|
sbp       Pooled          Equal          30      -1.40      0.1723
sbp       Satterthwaite   Unequal        30      -1.41      0.1680

                    Equality of Variances
Variable    Method      Num DF    Den DF    F Value    Pr > F
sbp         Folded F        16        14       1.39    0.5414

The REG Procedure
                    Analysis of Variance
                             Sum of         Mean
Source            DF        Squares       Square  F Value    Pr > F
Model              1      393.09816    393.09816     1.95    0.1723
Error             30     6032.87059    201.09569
Corrected Total   31     6425.96875

                        Parameter Estimates
                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|
Intercept     1      140.80000        3.66147      38.45      <.0001
smk           1        7.02353        5.02350       1.40      0.1723

                        Parameter Estimates
                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|
Intercept     1      144.31176        2.51175      57.45      <.0001
smk1          1        3.51176        2.51175       1.40      0.1723
```


SAS Results

- Same ANOVA table for each regression model
  - Same estimate of σ^2 and R^2
  - Identical Model and Error degrees of freedom
- Different slope and parameter estimates
- Same t-test for slope
- Using SMK $\begin{array}{ll} Y = \beta_0 + \beta_1(0) & \text{when SMK=0} \\ Y = \beta_0 + \beta_1(1) & \text{when SMK=1} \end{array}$

  this implies the intercept should equal the average SBP for smokers (SMK=0) and the slope should equal the difference in mean SBP. As you can see from the t-test results, this is the case.
- Using SMK1 $\begin{array}{ll} Y = \beta_0 + \beta_1(-1) & \text{when SMK1=-1} \\ Y = \beta_0 + \beta_1(1) & \text{when SMK1=1} \end{array}$

  Because the number of smokers/nonsmokers in the sample are not equal, the intercept does not equal the overall SBP average. Based on this model the sum of the Y's would equal $32\beta_0{+}2\beta_1$. In turn this means $\hat{\beta}_0 = \overline{Y}{-}2\hat{\beta}_1/32$, the "adjusted" grand mean (adjusted to account for more smokers in the sample). The slope is half as large because the change from nonsmokers to smokers is now 2 units of "X" (-1 to 1) rather than 1 (0 to 1).

Using Dummy Variable to Compare Regression Lines

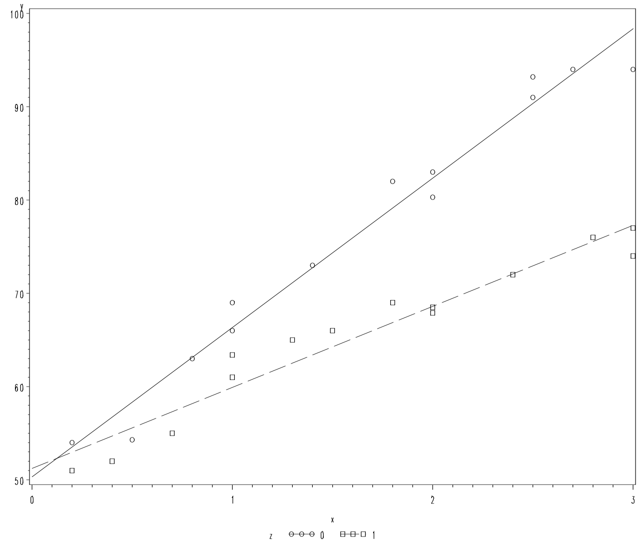
- When comparing regression lines, interested in
  - Are the slopes the same? (i.e., parallel lines)
  - Are the intercepts the same?
  - Are the lines the same?
- Can use dummy variables and interaction terms to test these hypotheses.
- If two groups, define

  $$Z = \begin{cases} 1 & \text{if obs from Grp 1} \\ 0 & \text{otherwise} \end{cases}$$

  and the interaction term $X\_Z = X \times Z$
- This approach is termed Method II in text
- Only appropriate when constant variance across groups
- Can modify Method I to handle non-constant variance across groups

Exercise 14-3

Comparing the relationship between age (AGE) and height (HGT) across two diet groups that are based on the protein content (RICH or POOR). The following scatterplot (with regression lines) summarizes the data.



The REG Procedure

PROTEIN RICH

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2484.01309	2484.01309	425.28	<.0001
Error	11	64.24999	5.84091		
Corrected Total	12	2548.26308			

Root MSE	2.41680	R-Square	0.9748
Dependent Mean	76.67692	Adj R-Sq	0.9725
Coeff Var	3.15192		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	50.32370	1.44303	34.87	<.0001
x	1	16.00897	0.77629	20.62	<.0001

PROTEIN POOR

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	856.70426	856.70426	186.34	<.0001
Error	12	55.17002	4.59750		
Corrected Total	13	911.87429			

Root MSE	2.14418	R-Square	0.9395
Dependent Mean	65.55714	Adj R-Sq	0.9345
Coeff Var	3.27070		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	51.22517	1.19612	42.83	<.0001
x	1	8.68604	0.63631	13.65	<.0001

From both the scatterplot and regression analysis, we see that the intercepts appear fairly close but the slope of the RICH group is much larger. In terms of the problem, this suggests that children grow faster if they have a protein rich diet.

As a check for constant variance, we can compare the MSE's from each individual model. In this case, the difference (4.60 vs 5.84) is small enough that the constant variance assumption appears reasonable.

If you did not want to make the constant variance assumption, here is a way to modify Method I to handle this. It is simply extending the ideas of to the two sample t-test with unequal variances. That formula is shown below.

$$T = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\text{Var}(\overline{Y}_1) + \text{Var}(\overline{Y}_2)}}$$

Comparing Slopes: Instead of computing a pooled mean square error and using this to determine the standard error of the difference in slopes (Method I), simply use the std errors of the slopes given in the SAS output. In this case, they are SE($\widehat{\beta}_{1R}$)=.77629 and SE($\widehat{\beta}_{1P}$)=.63631 so the standard error of the difference is

$$S_{\widehat{\beta}_{1R}-\widehat{\beta}_{1P}} = \sqrt{.77629^2 + .63631^2} = 1.00375$$

The test statistic is then

$$T = \frac{\widehat{\beta}_{1R} - \widehat{\beta}_{1P}}{S_{\widehat{\beta}_{1R}-\widehat{\beta}_{1P}}} = \frac{16.00897 - 8.68604}{1.00375} = 7.296$$

The degrees of freedom would be the smaller of $n_R - 2 = 11$ and $n_P - 2 = 12$. We'd reject in this case since the critical value is 2.201.

If we assume constant variance and pool, the pooled MSE is

$$\frac{11(5.84091) + 12(4.59750)}{11 + 12} = \frac{64.24999 + 55.17002}{23} = 5.1922$$

and the test statistic is

$$T = \frac{16.00897 - 8.68604}{\sqrt{5.1922\left(\frac{.77629^2}{5.84091} + \frac{.63631^2}{4.59750}\right)}} = 7.349$$

This would have degrees of freedom 23 and would also be significant. This standard error formula is similar to the text's except that I use only SAS output and the formula

$$\text{Var}(\widehat{\beta}_1) = \frac{\text{MSE}}{(n-1)S_X^2} \longrightarrow \frac{\text{Var}(\widehat{\beta}_1)}{\text{MSE}} = \frac{1}{(n-1)S_X^2}$$

This is the exact same test that you can get from the model with the dummy variable and interaction term.

SAS Commands

To utilize SAS, must create a dummy variable and interaction term. I create the variable z and x_z in the data step.

```
data problem3;
 infile 'i:\.www\datasets502\EX1403.DAT' firstobs=2 dlm='09'x;
 input country $ x y;
 if country = "RICH" then z=0;
 if country = "POOR" then z=1;
 x_z = x*z;

/* Generate the scatterplot */
proc sort; by z;
symbol1 v=circle i=rl;
symbol2 v=square i=rl line=2;
proc gplot;
 plot y*x = z;

/* Generate separate regressions for each level of z */
proc reg;
 model y = x;
 by z;

/* Generate the full regression model */
proc reg;
 model y = x z x_z / ss1;

run;
quit;
```


Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4174.20665	1391.40222	267.98	<.0001
Error	23	119.42001	5.19217		
Corrected Total	26	4293.62667			

Root MSE	2.27863	R-Square	0.9722
Dependent Mean	70.91111	Adj R-Sq	0.9686
Coeff Var	3.21337		

Parameter Estimates

Variable	DF	Parameter Estimate	Std Error	t Value	Pr > |t|	Type I SS
Intercept	1	50.32370	1.36053	36.99	<.0001	135766
x	1	16.00897	0.73192	21.87	<.0001	3053.34835
z	1	0.90148	1.86194	0.48	0.6329	840.45239
x_z	1	-7.32293	0.99647	-7.35	<.0001	280.40592

Understanding the Output

- For RICH : $z = 0$ and $x\_z = 0$ so the equation is

$$Y = 50.3237 + 16.00897X$$

This is the same line estimate as the individual output

- For POOR: $z = 1$ and $x\_z = x$ so the equation is

$$\begin{aligned} Y &= (50.3237 + 0.90148) + (16.00897 - 7.32293)X \\ &= 51.2252 + 8.68604X \end{aligned}$$

This is the same line estimate as the individual output

- Coefficient on z compares intercepts
- Coefficient on x_z compares slopes
- Both tests exactly the same as Method I (pooled)
- Can also test if lines equal (partial F test)

$$F = \frac{(840.45239 + 280.40592)/2}{5.19217} = 107.94$$

Reject since $F_{.05,2,23} = 3.42$.