Self-Assessment
Weeks 8: Multiple Regression with Qualitative Predictors; Multiple Comparisons

1. Suppose we wish to assess the impact of five treatments while blocking for study participant race (Black, Hispanic, White) on an outcome Y. How would these five treatments and three race categories be coded as dummy variables? Present actual data to illustrate the coding. Note: The term "blocking" used above represents analysis of variance language and indicates a variable for which one wishes to control statistically by including it in the model because the researcher believes it accounts for (predicts) variability in the outcome (DV). For example, when studying the effects of different fertilizers on tomato production, it is important to block other factors that can affect tomato production such as soil conditions and irrigation levels. Thus, blocking a variable simply means including it in the model so error variance can be reduced and thereby produce more powerful (i.e., lower Type 2 errors) statistical tests of the treatment.

2. Both the Bonferroni and Scheffé adjustments are designed to hold the familywise Type 1 error rate to a specific level. Can we be assured that both function to do this? One way to test this is to calculate the inflation to the Type 1 error rate using the adjusted Bonferroni and Scheffé per-comparison alpha (Type 1 error rate per test or per comparison).

The Week 6 Self-assessment Activity Question 4 asked that you calculate both Bonferroni and Scheffé confidence intervals for a study containing n = 76 observations with four drug treatments and a familywise error rate of .05.

DV = Heart rate = beats per minute
IV = Blood Pressure Medication = four drugs prescriptions (Losartan, Ziac, Lisinopril [12.5mg], and Lisinopril [40mg])

Critical t-values were as follows

Bonferroni adjusted critical t ratio
= ± 2.7129

Scheffé adjusted critical t ratio
= ±2.8627

As a review, the appendix below shows how these values were obtained.

Both of these critical t-ratios have corresponding specific pairwise comparison alpha levels. The alpha values have been adjusted using either the Bonferroni or Scheffé procedure.

To determine the corresponding specific alpha level for each, we can use Excel to find the two-tailed significance level and this will be the alpha level for each pairwise comparison.

Bonferroni adjusted pairwise alpha level
=T.DIST.2T(2.7129,72) = .008338

Scheffé adjusted pairwise alpha level
=T.DIST.2T(2.8627,72) = 0.005497

So these numbers tell us that if we wished to compare a p-value for each comparison against an alpha level, the Bonferroni adjusted alpha level would be .008338 and the Scheffé adjusted alpha level would be .005497.

(a) If one performed six pairwise comparisons using the Bonferroni adjusted pairwise comparison alpha of .008338, what would be the calculated familywise error rate across these six comparisons?

(b) If one performed six pairwise comparisons using the Bonferroni adjusted pairwise comparison alpha of .005497, what would be the calculated familywise error rate across these six comparisons?

3. Ian Walker collected data on bicycle overtaking (vehicles passing bicycles) in the UK. His data are available from this link:

http://drianwalker.com/overtaking/

For this self-assessment activity we will focus on the following variables:

Dependent Variable
Passing_distance = distance in meters that vehicles gave bicycles while passing

Predictor Variables
Distance_from_kerb = distance of bicycle from curb. Distances, in meters, were 0.25, 0.50, 0.75, 1.00, and 1.25.
Helmet = whether rider used a helmet (1 = yes, 0 = no)
Car = whether vehicle that passed was a car or some other vehicle type (e.g., bus, lorry, etc.; 1 = car, 0 = other)
Time = time of day overtaking recorded grouped into three categories, morning, midday, and afternoon

Two of the above variables contain more than two categories, so dummy variables were constructed as follows:

Distance_from_kerb dummy variables:

- Curb_0.25 (1 = yes, 0 = no)
- Curb_0.50 (1 = yes, 0 = no)
- Curb_0.75 (1 = yes, 0 = no)
- Curb_1.00 (1 = yes, 0 = no)
- Curb_1.25 (1 = yes, 0 = no)

Time dummy variables:

Morning	(1 = yes, 0 = no; this represents times of 7am to 10:59am)
Midday	(1 = yes, 0 = no; from 11am to 2pm)
Afternoon	(1 = yes, 0 = no; between 2:01pm and 6pm)

Below is an SPSS regression analysis of passing_distance regressed on the four predictors outlined above.

Descriptive Statistics

	Mean	Std. Deviation	N
Passing distance	1.31391	.383454	2355
Car	.7253	.44648	2355
helmet	.49	.500	2355
Curb_0.50	.2314	.42183	2355
Curb_0.75	.1439	.35111	2355
Curb_1.00	.1992	.39945	2355
Curb_1.25	.1410	.34807	2355
Midday	.2917	.45465	2355
Afternoon	.3125	.46362	2355

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.285(a)	.081	.078	.368225	.081	25.844	8	2346	.000

a Predictors: (Constant), Afternoon, Curb_0.50, Car, helmet, Curb_1.25, Curb_0.75, Midday, Curb_1.00

ANOVA(c)

Model			Sum of Squares	df	Mean Square	F	Sig.	R Square Change
1	Subset Tests	Car	2.368	1	2.368	17.465	.000(a)	.007
		helmet	1.806	1	1.806	13.322	.000(a)	.005
		Curb_0.50, Curb_0.75, Curb_1.00, Curb_1.25	14.874	4	3.719	27.426	.000(a)	.043
		Midday, Afternoon	.022	2	.011	.081	.922(a)	.000
	Regression		28.033	8	3.504	25.844	.000(b)	
	Residual		318.093	2346	.136			
	Total		346.126	2354				

a Tested against the full model.
b Predictors in the Full Model: (Constant), Afternoon, Curb_0.50, Car, helmet, Curb_1.25, Curb_0.75, Midday, Curb_1.00.
c Dependent Variable: Passing distance

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1.406	.028		49.808	.000	1.350	1.461
	Car	.072	.017	.083	4.179	.000	.038	.105
	helmet	-.057	.016	-.074	-3.650	.000	-.087	-.026
	Curb_0.50	-.092	.023	-.101	-4.028	.000	-.137	-.047
	Curb_0.75	-.173	.028	-.158	-6.163	.000	-.228	-.118
	Curb_1.00	-.184	.026	-.191	-7.083	.000	-.235	-.133
	Curb_1.25	-.267	.027	-.242	-9.804	.000	-.320	-.214
	Midday	.008	.021	.009	.386	.700	-.032	.048
	Afternoon	.007	.022	.008	.298	.765	-.037	.050

a Dependent Variable: Passing distance

(a) Note that all predictor variables included in this regression are dummy variables with 0, 1 coding. The "Descriptive Statistics" table shows the following means:

Variable	Mean
Car	.7253
helmet	.4900
Curb_0.50	.2314
Curb_0.75	.1439
Curb_1.00	.1992
Curb_1.25	.1410
Midday	.2917
Afternoon	.3125

(a1) What does the mean value of .7253 for Car tell us? What is the interpretation of this value?

(a2) For helmet, the mean is .4900, what does this tell us?

(a3) For Curb_0.75 the mean value is .1439, what does this tell us?

(a4) For midday the mean value is .2917 – what interpretation may we use for this?

(b) The squared semi-partial correlations (ΔR^2) for each of the predictors are

Car Type	= 0.007
Helmet Use	= 0.005
Curb Distance	= 0.043
Time of Data	= 0.000

When added together, these produce a summed R^2 value of 0.007 + 0.005 + 0.043 + 0.000 = 0.055. However, SPSS reports that the total model R^2 is .081. Why is there a discrepancy between the summed R^2 and the model R^2 reported by SPSS?

(c) The ANOVA table shows us F ratios and p-values for each predictor variable. For Helmet use, F = 13.322 and p = .000, so there are differences in passing distances between riders wearing helmets and riders not wearing helmets. Suppose for a moment that the ANOVA table was not presented so we don't have access to this F ratio or ΔR^2 values. Would we be able to determine whether the null for helmet use could be rejected with any other information provided in the regression output? If yes, what information could we use?

(d) As noted above in (c), the ANOVA table shows us F ratios and p-values for each predictor variable. For Curb (Kerb in the UK) Distance, F = 27.426 with p = .000. Suppose for a moment that the ANOVA table was not presented so we don't have access to this F ratio or ΔR^2 values. Would we be able to determine whether the global test of the null for Curb Distance could be rejected with any other information provided in the regression output? If yes, what information could we use?

(e) Provide literal interpretations for each of the unstandardized regression coefficients listed below.

Intercept, B0 = 1.406:
Car, B1 = .072:
Helmet, B2 = -.057:
Curb_1.00, B5 = -.184:
Afternoon, B8 = .007:

(f) What is the predicted mean passing distance for someone with the following variable values:

Scenario 1:
Passing Car
Not wearing a helmet
Curb distance of .25
Midday riding

Scenario 2:
Passing Truck
Wearing a helmet
Curb distance of 1.25
Morning Riding

(g) Which factors (predictors) are not statistically associated with passing distance?

(h) What is the interpretation for the 95% confidence interval for b4 (Curb 0.75 dummy)?

(i) Suppose one wished to perform all pairwise comparisons among curb distances and also among time of day. Using the Bonferroni correction, what would be the adjusted Bonferroni alpha (Type 1 error rate) per comparison if the familywise error rate is to be .05?

(j) Which of the four independent variables appears to be the strongest predictor of passing distances?

4. Below is a data file containing the following variables for cars taken between 1970 and 1982:

mpg:	miles per gallon
engine:	engine displacement in cubic inches
horse:	horsepower
weight:	vehicle weight in pounds
accel:	time to accelerate from 0 to 60 mph in seconds
year:	model year (70 = 1970, to 82 = 1982)
origin:	country of origin (1=American, 2=Europe, 3=Japan)
cylinder:	number of cylinders

SPSS Data: http://www.bwgriffin.com/gsu/courses/edur8132/selfassessments/Week04/cars_missing_deleted.sav
(Note: There are underscore marks between words in the SPSS data file name.)
Other Data Format: If you prefer a data file format other than SPSS, let me know.

For this problem we wish to know whether MPG differs among car origins and number of cylinders:

Predicted MPG = b0 + origin of car with appropriate dummy variables + number of cylinders

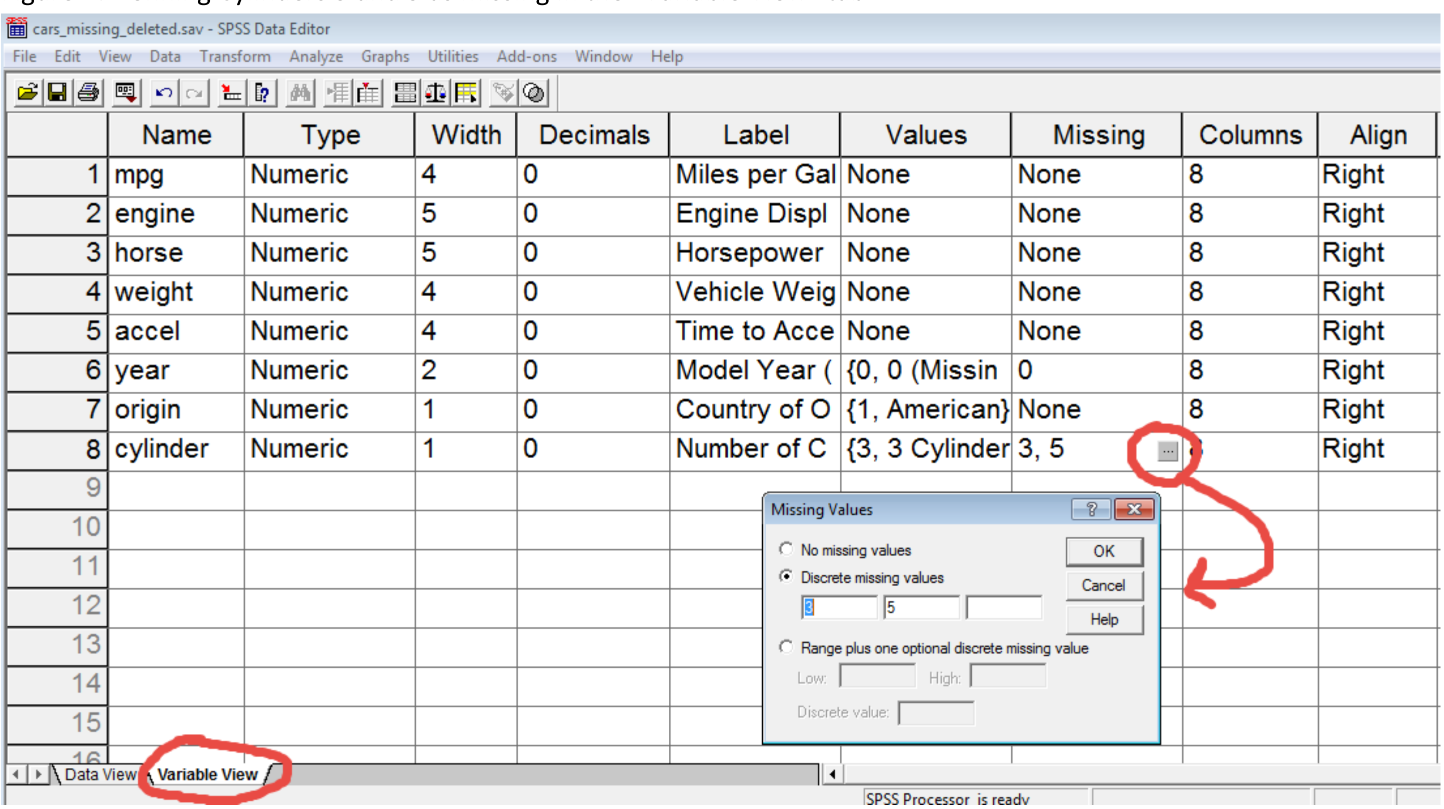
Origin of car is categorical. Number of cylinders may appear to be ratio, but since observed categories of this variable ares limited, it is best to treat this variable as categorical. Note the following number of cylinders reported:

Number of Cylinders

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3 Cylinders	4	1.0	1.0	1.0
	4 Cylinders	199	50.9	50.9	51.9
	5 Cylinders	3	.8	.8	52.7
	6 Cylinders	83	21.2	21.2	73.9
	8 Cylinders	102	26.1	26.1	100.0
	Total	391	100.0	100.0	

As the frequency display above shows, the number of cylinders include 3, 4, 5, 6, and 8. However, only 4 cars had 3 cylinders and only 3 cars had 5 cylinders. Given the small sample sizes for these categories, it is best to remove these cases from the regression analysis. There are several ways to accomplish this. Three approaches are (a) manually delete these cases after sorting all cases on number of cylinders, (b) telling SPSS to treat these 7 cases as missing values so they will not be included in any analysis (use Recode into Same Variable and set 3 Cylinders and 5 Cylinders as system missing), or (c) defining 3 and 5 Cylinders as missing values in the variable missing values (see Figure 1 below for how this is accomplished in SPSS). Other possibilities also exist.

Figure 1: Defining Cylinders 3 and 5 as missing in the “Variable View” tab

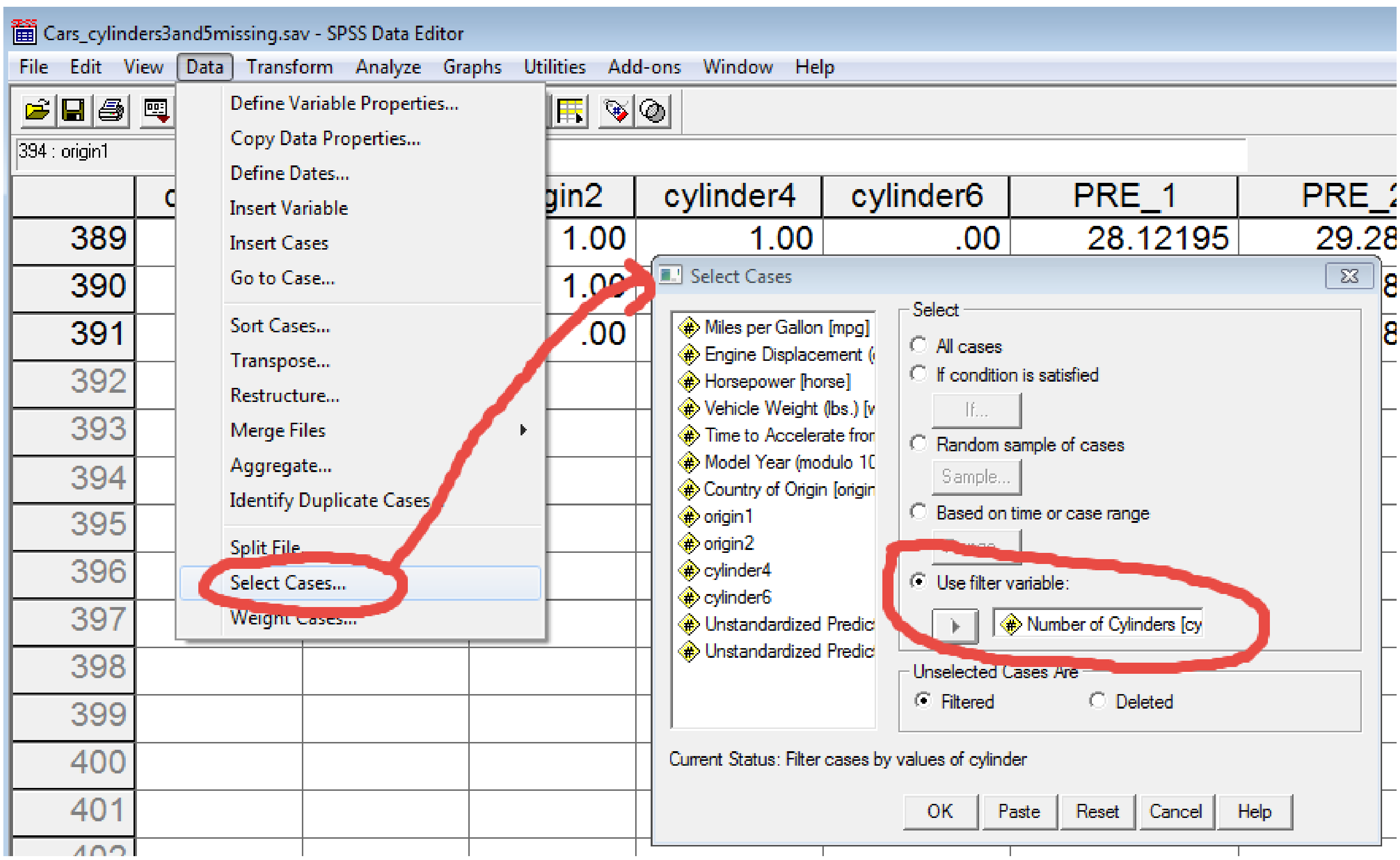


After defining Cylinders 3 and 5 as missing as illustrated in the Figure 1 above, I re-ran the Frequency command for Cylinders and obtained the following results. Note that Cylinders 3 and 5 are now identified as missing and SPSS will automatically discard these cases when performing various statistical tests IF the variable Cylinders is used in the analysis. If you use dummy variables created from Cylinders, then you need to tell SPSS to select only those cases that are complete for Cylinders. Use the Select Cases command as illustrated in Figure 2 below and identify the variable Cylinders as the selection filter variable. This tells SPSS to only use cases with complete Cylinder information – missing cases are ignored in all analyses.

Number of Cylinders

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	4 Cylinders	199	50.9	51.8	51.8
	6 Cylinders	83	21.2	21.6	73.4
	8 Cylinders	102	26.1	26.6	100.0
	Total	384	98.2	100.0	
Missing	3 Cylinders	4	1.0		
	5 Cylinders	3	.8		
	Total	7	1.8		
Total		391	100.0		

Figure 2: Select only those cases with complete Cylinders data



Present an APA styled regression analysis with DV = MPG, IV = origin, and IV = Cylinders (4, 6, and 8 only). Set alpha = .01. You will have to create the dummy variables for origins and Cylinders. Also present Scheffé confidence intervals comparisons among origins and among cylinders.

Appendix

Question 2

Review: Determining Bonferroni and Scheffé critical t values for confidence interval construction.

Study consisted of n = 72 observations on heart rate across four medications.

DV = Heart rate = beats per minute
IV = Blood Pressure Medication = four drugs prescriptions (Losartan, Ziac, Lisinopril [12.5mg], and Lisinopril [40mg])

Wish to maintain a familywise error rate of .05.

Bonferroni adjusted critical t ratio:

(a) Adjusted alpha per comparison is .05/6 = .083333 (divide by 6, the number of possible pairwise comparisons among four drug treatments)
(b) Study degrees of freedom is n – k – 1 where k is the number of dummy variables (number of groups minus 1), so 76 – 3 – 1 = 72
(c) Then use Excel critical t function to find the critical t-value:

=T.INV.2T(adjusted alpha, df)
=T.INV.2T(.008333, 72)
= ± 2.7129

Scheffé adjusted critical t ratio:

(a) Since the Scheffé adjusted critical t is based upon an F ratio, we must determine the critical F by first finding the model degrees of freedom

df1 = J – 1 = 4 – 1 = 3
df2 = n – k – 1 = 76 – 3 – 1 = 72

where J is the number of groups, and k is the number of dummy variables in the regression equation.

(b) Next find the critical F ratio for a familywise error rate of .05. This can be found using Excel

=F.INV.RT(alpha level, df1, df2)
=F.INV.RT(0.05,3,72)
= 2.7318

(c) Next convert this critical F ratio to a Scheffé adjusted F ratio

Scheffé F = (J – 1) (original critical F)
Scheffé F = (3) (2.7318)
Scheffé F = 8.1954

(d) Next convert this Scheffé F ratio to a critical Scheffé t value by taking the square root of the Scheffé F:

Scheffé t = $\sqrt{\text{Scheffé F}}$

Scheffé t = $\sqrt{8.1954}$

Scheffé t = ±2.8627

Now we have the critical t-value used to test the six possible pairwise comparisons among four drug treatments with an overall familywise error rate of .05 or less.