

Stat 102

Introduction to Business Statistics

Class 18

Richard P. Waterman

Wharton

Table of contents I

1 Today's module

2 Last time

3 More on multiplicity

4 Summary

5 Next time

Today's module

Topics to be covered in this module:

- Last time
- Ways of correcting for multiplicity:
 - 1 Tukey's HSD
 - 2 Bonferroni
 - 3 The False Discovery Rate, adjusting the p-values
- Summary
- Next time

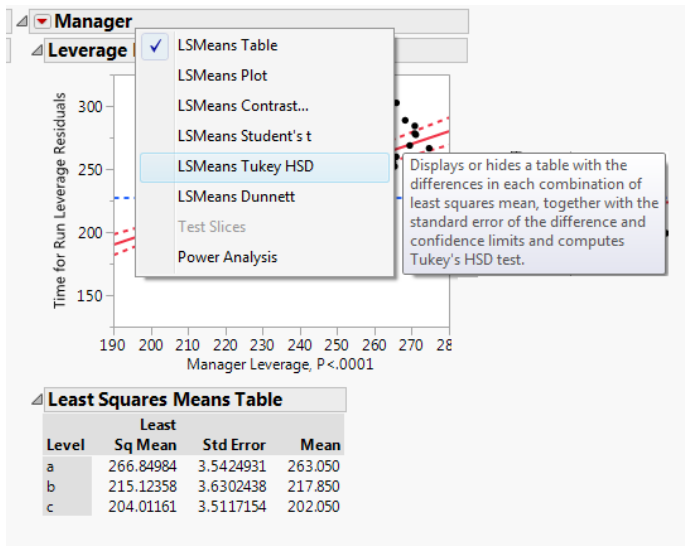
- Working out the line for each group in a categorical variable regression
- The Least Squares Means table
- Multi-level categoricals
- Testing for multi-level categoricals

- Tukey's HSD is a technique for pairwise comparisons, that controls the *familywise error rate*.
- The *use case*: you have a large multi-level categorical variable in a regression and you want to compare the levels.
 - Which levels are significantly different from each other (in terms of the mean of y) controlling for the other x 's?
 - Which levels might you collapse together, to make a more parsimonious representation of the categorical variable?

The Tukey HSD procedure

- The relevant JMP output is found by drilling down on the title bar of the leverage plot for the categorical variable.
- HSD stands for *Honest Significant Difference*.
- I will start with the ProdTime data set, then do a more challenging example.

The Tukey HSD procedure



Output from the Tukey HSD procedure

Least Squares Means Table

Level	Sq Mean	Std Error	Mean
a	266.84984	3.5424931	263.050
b	215.12358	3.6302438	217.850
c	204.01161	3.5117154	202.050

LSMeans Differences Tukey HSD

$\alpha = 0.050$ $Q = 2.40999$

		LSMean[j]		
Mean[i]-Mean[j]		a	b	c
Std Err Dif				
Lower CL Dif				
Upper CL Dif				
a	LSMean[i]	0	51.7263	62.8382
		0	5.07227	4.98813
		0	39.5022	50.8169
		0	63.9504	74.8596
b	LSMean[i]	-51.726	0	11.112
		5.07227	0	5.05082
		-63.95	0	-1.0605
		-39.502	0	23.2844
c	LSMean[i]	-62.838	-11.112	0
		4.98813	5.05082	0
		-74.86	-23.284	0
		-50.817	1.06045	0

Level		Sq Mean
a	A	266.84984
b	B	215.12358
c	B	204.01161

Levels not connected by same letter are significantly different.

Discussion

- The table shows all the pairwise comparisons between the levels of the categorical variable.
- There are four numbers in each cell:
 - ① The difference in least squares means between the two groups
 - ② The standard error of the difference in means
 - ③ The lower end of the 95% CI for the difference in means
 - ④ The upper end of the 95% CI for the difference in means
- If the confidence interval does not contain 0 then the cell is colored red, to indicate a significant difference.
- The key feature of the table is how the confidence intervals are calculated.
- They do **not** use $\pm 2Std.Err.$
- They use the “Q” number at the top of the output instead (here 2.40999) to calculate the interval.

- The adjustment for multiplicity comes by replacing the usual **2** with the Q-value, here 2.40999.
- It makes the individual confidence intervals wider and therefore harder to reject 0 being a feasible value for the difference in means.
- The adjustment recognizes that there are three possible pairwise comparisons and controls the familywise error rate at $\alpha = 0.05$ over all the comparisons.

Using Tukey in an exploratory way

- Consider the PRSM data set with the 30+level categorical, SIC code. We'll look at SIC code in a multiple regression via Tukey's HSD and see what we learn:
- To do in class.
- Looking at the p-values on the t-stats is risky as there are so many of them and we may well be experiencing a Type I error.
- Looking at the Tukey HSD output is more principled as it recognizes all the comparisons we are making, when there is no a-priori hypothesis set.
- If you are going *data snooping* use a procedure, such as Tukey's, to make it more legitimate.

- Over last 25 years we have had:
 - Access to increasing quantities of data.
 - Access to increased computing power.
- Consequently we may find ourselves in a position where we want to test hundreds or even thousands of hypotheses.
- For example a typical microarray experiment might result in testing 10,000 hypotheses.
- This gives rise to issues that are novel and not always addressed by analysts. The essential problem is termed “multiplicity”.

More on multiplicity

- Multiplicity refers to the situation where we have no upfront specific hypothesis and are essentially on a fishing expedition.
- The problem is that every hypothesis test has the potential for error. Typically we control the Type I error rate (false positive) at $\alpha = 0.05$. But the more hypothesis tests you do, the more Type I errors you will make.
- If we ran 10 independent hypothesis tests, each at $\alpha = 0.05$ then the probability that one or more Type I errors are made given all null hypotheses are true is: $1 - (1 - 0.05)^{10} = 0.4$.
- If we ran 100 independent tests then this probability increases to 0.994. So we will almost certainly have a number of false positives on our hands.
- The probability of one or more Type I errors is called the familywise error rate (FWER). We can see from the above calculations that it can get very large.

- A popular and very simple way to control the FWER is to use what is called the Bonferroni correction.
- If you are performing k tests, then rather than use a p-value cut-off of α , you would use α/k as the cut-off value.
- If you use this adjusted cut-off value then it can be shown that the FWER must be less than α .
- Example: You are going to do 100 hypothesis tests. Bonferroni says don't use a p-value cut-off of 0.05, but rather use $0.05/100 = 0.0005$.
- That is, make it much harder to find statistical significance. This is just what Tukey did in the Tukey HSD procedure by making the confidence intervals wider.

The False Discovery Rate (FDR)

- There is a drawback to the simple Bonferroni procedure. It is very conservative.
- When there are real effects to be found (the alternative hypothesis is true) you may well miss them because the p-value cut-off is so stringent.
- A modern approach to multiplicity is to accept that you are going to do many hypothesis tests upfront and be willing to accept some Type I errors.

The False Discovery Rate (FDR)

- The following table summarizes the possible outcomes of k hypothesis tests.

	Declared not significant	Declared significant (discovery)	Total
Null true	U	V	k_0
Alternative true	T	S	$k - k_0$
Totals	$k - R$	R	k

- In this table V denotes the number of **false discoveries**. The False Discovery Rate (FDR) is defined as $E(V/R)$ and the goal is to keep this number below some set value, sometimes referred to as q .

The False Discovery Rate (FDR)

- JMP now has an implementation of an FDR procedure in the multiple regression platform.
- It adjusts the p-values upwards to make it harder to reject the null hypothesis.
- If you use a cut-off of 0.05 on the adjusted p-values (sometimes called q-values) then you can be sure that the FDR is less than 0.05.
- All of the Tukey-HSD, Bonferroni correction and the FDR address multiplicity.
- They do so by making it harder to declare an effect significant as compared to using the naïve p-value.

The False Discovery Rate Example

- The dataset `StockFishing.jmp` contains 4 years of returns on the SPY ETF which tracks the value of the S&P500.
- It also has almost 100 explanatory variables derived from technical indicators for 12 different stocks.
- If we fit a multiple regression model using all of the predictors for the next day's return on SPY we find that 4 of the variables are indicated as significant at the $\alpha = 0.05$ level as shown in the summary of the p-values.
- However, with about 100 predictors these may well be “false discoveries”. Using the FDR adjusted p-values ¹we get a different conclusion.
- After adjusting the p-values, none of them are significant which makes more sense given the belief that it is almost impossible to predict short-term returns on the stock market.

¹To obtain the FDR adjusted p-values click on the FDR checkbox in the effect summary table in the multiple regression output

The False Discovery Rate Example

Figure 1: Unadjusted p-values






Source	LogWorth		PValue
XOM Volume	1.976		0.01056
NVDA Low	1.976		0.01058
SPY High	1.965		0.01084
AMZN High	1.413		0.03866
JPM Open	1.296		0.05061

Figure 2: FDR adjusted p-values (aka q-values)

Source	FDR LogWorth	FDR PValue
XOM Volume	0.523	0.29981
NVDA Low	0.523	0.29981
SPY High	0.523	0.29981
AMZN High	0.216	0.60832
JPM Open	0.216	0.60832

Summary

- Multiplicity and how to correct for it:
- Tukey's HSD
- Bonferroni
- The False Discovery Rate

Next time

- Introduction to step-wise regression