Predicting Attendance

The concessions manager at a professional indoor soccer team's stadium needs to be able to predict attendance several days before home games (in order to make staffing decisions), and to refine those predictions on the morning of game day (when food must be defrosted and sent from off-site to the stadium).

The quality of the incoming opponent (measured by that opponent's recent winning percentage), as well as the amount (in dollars) of advertising planned by the team's owners over the days leading up to a game, are known to him well in advance of game day. On the morning of game day, he can, of course, observe whether it's clear ("weather"=0) or raining ("weather"=1).

All home games are played on either Friday evening, or Saturday or Sunday afternoon. Data collected over the past 40 home games is available here, and some of it is exhibited below:

	A	B	C	D	E	F
1	**Attend**	**Quality**	**Advert**	**Weather**		**Game day**
2	16935	0.6188	3053.56	1		Sunday
3	12225	0.0891	3356.33	0		Friday
4	15394	0.8699	2846.98	0		Saturday
5	14518	0.5458	3093.49	0		Saturday
6	15044	0.5677	3353.32	0		Sunday
7	16395	0.5891	3353.61	1		Saturday
8	14891	0.6590	2860.94	0		Sunday
9	16885	0.6376	2843.72	1		Sunday
10	13722	0.2873	3117.05	0		Sunday
11	...					

Ignoring game day for the moment, regressing attendance onto the first three explanatory variables yields these results:

	A	B	C	D	E	F
1		**Regression: Attend**				
2			**constant**	**Quality**	**Advert**	**Weather**
3		**coefficient**	10278.5986	4333.25952	0.53656238	1554.02007
4		**std error of coef**	1259.23753	553.754552	0.34196535	182.378164
5		**t-ratio**	8.1626	7.8252	1.5691	8.5209
6		**significance**	0.0000%	0.0000%	12.5384%	0.0000%
7		**beta-weight**		0.7258	0.1461	0.6679
8						
9		**standard error of regression**		554.019822		
10		**coefficient of determination**		78.74%		
11		**adjusted coef of determination**		76.97%		

Were there no other variables to consider, this would suggest that fans like to see good opponents, and that rain brings more fans to the stadium (perhaps because it brings in those with planned-but-now-cancelled outdoor activities, seeking an indoor event to attend). As well, it appears that advertising does fill additional seats. However, the correlation matrix shows a curious phenomenon:

	A	B	C	D	E	F
1		Correlations				
2						
3			Attend	Quality	Advert	Weather
4		Attend	1.00000	0.59888	-0.32542	0.59940
5		Quality	0.59888	1.00000	-0.54346	-0.07107
6		Advert	-0.32542	-0.54346	1.00000	-0.11549
7		Weather	0.59940	-0.07107	-0.11549	1.00000

The correlation between advertising and attendance is negative!

[Backstory: I made up this dataset in the spring of 1998. The Chicago Bulls were heading for their sixth NBA championship, and many expected that Michael Jordan would retire at the end of the season. I read an article one day which noted that the Bulls' road games were all sell-outs: Fans in other cities were turning out in droves for what could be their last chance ever to see Jordan play. I thought to myself, "Sensible owners, expecting free advertising from local sportswriters and sportscasters and loads of fan excitement, won't waste money advertising when the Bulls come to town. Instead they'll spend their media budgets advertising the less-exciting home games against other opponents. The advertising will bring additional fans to the stadium, but won't generate sellouts -- and as a result, advertising and attendance could end up being negatively correlated." I chose to capture this in an exam dataset.]

The phenomenon is explained (without need to know the backstory) by considering the correlations between the explanatory variables: Apparently, the most advertising is being done against low-quality teams, and (from the regression) low-quality teams don't draw many fans to the stadium. Advertising helps (again, from the regression), but not enough to fully offset the low-quality-induced reduction in attendance.

How might we bring game-day into the regression model? By selecting one of the three possibly game days as our "foundation" case, and creating two dummy variables which indicate whether a game is played on one of the two non-foundation days.

Taking Friday as the foundation case, we create "Is the game played on Saturday?" and "Is the game played on Sunday?" variables. One way to do this in a spreadsheet is illustrated below:

	A	B	C	D	E	F	G	H
1	Attend	Quality	Advert	Weather	Saturday	Sunday		Game day
2	16935	0.6188	3053.56	1	0	1		Sunday
3	12225	0.0891	3356.33	0	0	0		Friday
4	15394	0.8699	[illegible]	[illegible]	1	0		Saturday
5	14518	0.5458	[illegible]	[illegible]	1	0		Saturday
6	15044	0.5677	[illegible]	[illegible]	0	1		Sunday
7	16395	0.5891	[illegible]	[illegible]	1	0		Saturday
8	14891	0.6590	2860.94	0	0	1		Sunday
9	16885	0.6376	2843.72	1	0	1		Sunday
10	13722	0.2873	3117.05	0	0	1		Sunday
11	...							

=IF(E1=H2,1,0)

=IF(F1=H2,1,0)

Here's the resulting regression:

	A	B	C	D	E	F	G	H
1		Regression: Attend						
2			constant	Quality	Advert	Weather	Saturday	Sunday
3		coefficient	8789.40904	4503.83617	0.79081721	1673.24942	622.745696	1150.75849
4		std error of coef	734.416718	298.686346	0.19572698	98.7554497	120.826634	120.607761
5		t-ratio	11.9679	15.0788	4.0404	16.9434	5.1540	9.5413
6		significance	0.0000%	0.0000%	0.0288%	0.0000%	0.0011%	0.0000%
7		beta-weight		0.7543	0.2154	0.7191	0.2645	0.4626
8								
9		standard error of regression		297.192323				
10		coefficient of determination		94.22%				
11		adjusted coef of determination		93.37%				

As the predictions below show, the pure game-day-related difference between playing on a Friday, Saturday, or Sunday is that Saturdays draw 622.75 more attendees (on average) than do Fridays, given the same opponent quality, advertising, and weather. And Sundays outdraw Fridays by 1150.76 (and consequently outdraw Saturdays by 528.01.

	A	B	C	D	E	F
1		Predictions, using most-recent regression				
2						
3		Predict	coefficients	values for prediction		
4		constant	8789.409			
5		Quality	4503.8362	0.6	0.6	0.6
6		Advert	0.7908172	300	300	300
7		Weather	1673.2494	1	1	1
8		Saturday	622.7457	0	1	0
9		Sunday	1150.7585	0	0	1
10						
11		predicted value of Attend		13402.21	14024.95	14552.96
12		standard error of prediction		619.6775	583.9528	608.1109
13		standard error of regression		297.1923	297.1923	297.1923
14		standard error of estimated mean		543.7619	502.6705	530.5427
15						
16		confidence level	95.00%			
17		t-statistic	2.0322			

In order to confirm that our encoding of game day didn't bias our results, let's recode, with Saturday as the foundation case:

	A	B	C	D	E	F	G	H
1	Attend	Quality	Advert	Weather	Friday	Sunday		Game day
2	16935	0.6188	3053.56	1	0	1		Sunday
3	12225	0.0891	3356.33	0	1	0		Friday
4	15394	0.8699	[illegible]	[illegible]	0	0		Saturday
5	14518	0.5458	[illegible]	0	0	0		Saturday
6	15044	0.5677	[illegible]	[illegible]	0	1		Sunday
7	16395	0.5891	[illegible]	[illegible]	0	0		Saturday
8	14891	0.6590	2860.94	0	0	1		Sunday
9	16885	0.6376	2843.72	1	0	1		Sunday
10	13722	0.2873	3117.05	0	0	1		Sunday
11	...							

=IF(E1=H2,1,0)

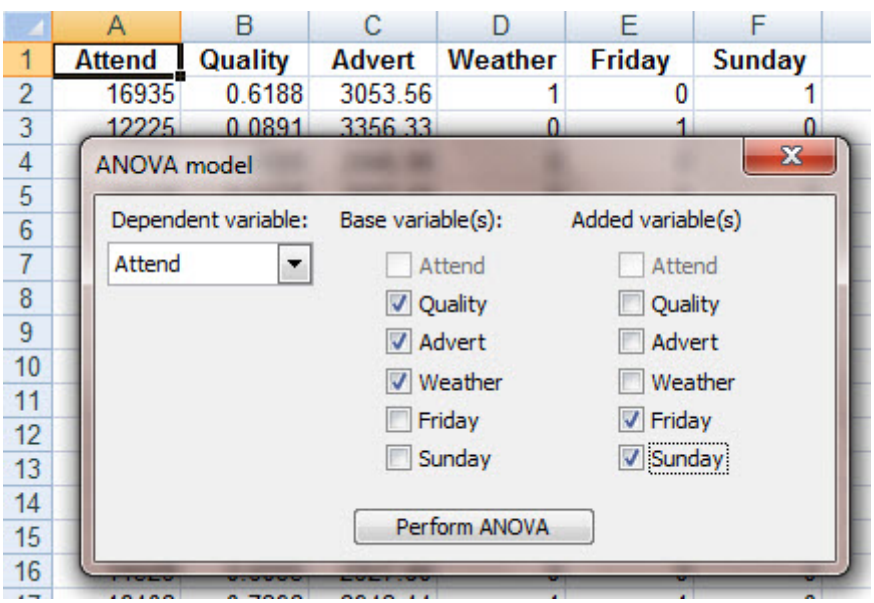
=IF(F1=H2,1,0)

In the new regression, the coefficients of the dummy variables are rearranged, but ultimately convey precisely the same message as before.

	A	B	C	D	E	F	G	H
1		Regression: Attend						
2			constant	Quality	Advert	Weather	Friday	Sunday
3		coefficient	9412.15473	4503.83617	0.79081721	1673.24942	-622.7457	528.012791
4		std error of coef	695.92694	298.686346	0.19572698	98.7554497	120.826634	118.694581
5		t-ratio	13.5246	15.0788	4.0404	16.9434	-5.1540	4.4485
6		significance	0.0000%	0.0000%	0.0288%	0.0000%	0.0011%	0.0088%
7		beta-weight		0.7543	0.2154	0.7191	-0.2559	0.2123
8								
9		standard error of regression		297.192323				
10		coefficient of determination		94.22%				
11		adjusted coef of determination		93.37%				

When we play this modeling trick, using a collection of dummy variables to represent one qualitative variable, one new issue arises: The significance levels of the individual dummy variables don't really tell us whether there's evidence (or not) supporting the inclusion of the qualitative variable (i.e., *all* of the dummy variables) in our model. If some of the individual significance levels are near 0, and others are large, what conclusion can we draw?

Fortunately, a technique known as "analysis of variance" (abbreviated ANOVA) answers this question. Imagine the variables in the model separated into two groups: the base group (of everything other than the dummy variables representing one qualitative variable), and the added group of dummy variables.

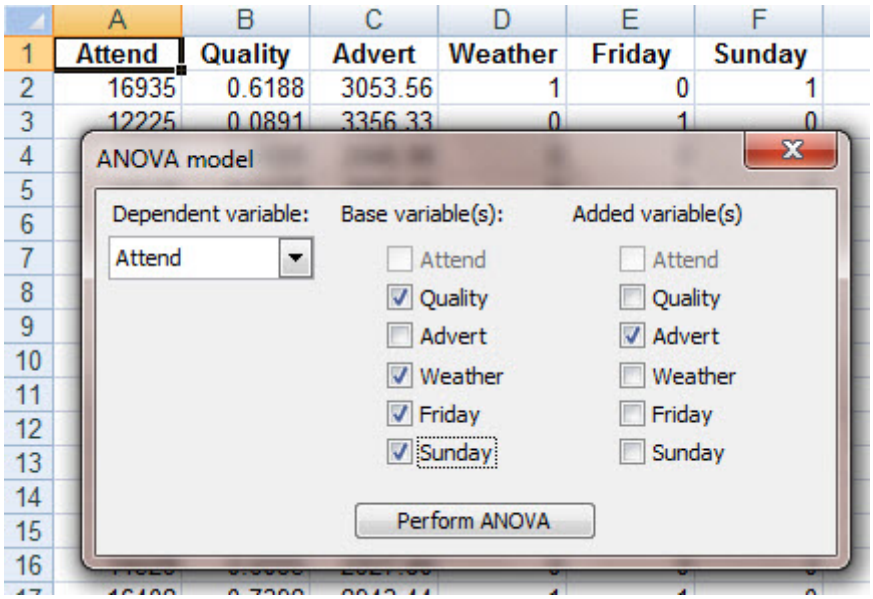


Hypothesize that the true coefficients of all of the added variables are 0, i.e., there's nothing amongst them that belongs in the model. Then the significance level of the sample data, with respect to this null hypothesis, is found at the lower right-hand corner of the ANOVA table (cell G10 on the ANOVA tab in KStat).

	A	B	C	D	E	F	G	H
1		Analysis of variance						
2			base model		extended model		difference	
3			sum of squares	df	sum of squares	df	sum of squares	df
4		regression	40927473.32	3	48974248.57	5	8046775.251	2
5		residual	11049766.66	36	3002991.405	34	3002991.405	34
6		total	51977239.98	39	51977239.98	39	11049766.66	36
7								
8		F-ratio	44.4471		110.8977		45.5530	
9		degrees of freedom	(3, 36)		(5, 34)		(2, 34)	
10		significance	0.00000%		0.00000%		0.00000%	

In the current case, that significance level is so close to zero that its first four decimal places are all 0, and we therefore conclude that the data overwhelmingly contradicts the null hypothesis, and supports the alternative statement that game day *does* belong in our model.

ANOVA directly generalizes the significance level we've previously looked at in the regression output. For example, using the ANOVA option to ask whether there's evidence that advertising belongs in our final model,



we obtain:

	A	B	C	D	E	F	G	H
1		Analysis of variance						
2			base model		extended model		difference	
3			sum of squares	df	sum of squares	df	sum of squares	df
4		regression	47532378.88	4	48974248.57	5	1441869.694	1
5		residual	4444861.1	35	3002991.405	34	3002991.405	34
6		total	51977239.98	39	51977239.98	39	4444861.1	35
7								
8		F-ratio	93.5706		110.8977		16.3249	
9		degrees of freedom	(4, 35)		(5, 34)		(1, 34)	
10		significance	0.00000%		0.00000%		0.02884%	

i.e., precisely the same significance level of 0.0288(4)% that we did in the regression.