

There are four major questions on this exam. Each sub-question is worth three points, unless otherwise indicated.

**Question A**

The output on the following pages examines predictors of the winning race time in the New York City Marathon over the last 21 years. The following variables are included:

TIME	Winning marathon time (in minutes)
YEAR	Year of the race (1978-1998)
TEM	High temperature in New York City on race day (in degrees Fahrenheit)
HUM	Relative humidity on race day (in percentage of saturation)

From these variables, some additional ones have been computed:

TEM <sub>D</sub>	TEM - 63.0476 (i.e, minus the mean TEM from all 21 years)
HUM <sub>D</sub>	HUM - 0.5333 (i.e., minus the mean HUM from all 21 years)
TEM <sup>2</sup>	TEM * TEM
TEM <sub>D</sub> <sup>2</sup>	TEM <sub>D</sub> * TEM <sub>D</sub>
TEMHUM	TEM* HUM
TEM <sub>D</sub> HUM <sub>D</sub>	TEM <sub>D</sub> * HUM <sub>D</sub>

Based on this output, answer the following questions.

1. Provide interpretations for the slope and the intercept in the first model, where YEAR is used to predict TIME.
2. Overall (ignoring HUM) is there a linear relationship between the high temperature on race day and the winning race time? (Provide PRE, F\*, and a substantive interpretation.)
3. Is there evidence that the relationship between the high temperature on race day and the winning race time is nonlinear? (Provide PRE, F\*, and a substantive interpretation concerning the nature of the nonlinearity.)
4. Provide a substantive interpretation for the parameter estimate associated with TEM<sup>2</sup> in Model 3.

5. Given the nonlinear relationship between the high temperature on race day and the winning time, is there a significant impact of temperature at the mean temperature level? (Provide PRE, F\*, and a substantive interpretation.)
6. Given the nonlinearity of the temperature - race time relationship, what is the best estimate of the optimal temperature for winning racers? (Hint: At cold temperatures, there seems to be a negative relationship, while at warmer temperatures, there is a positive relationship. At what temperature does the predicted race time reach a minimum?.)
7. Does the linear relationship between the high temperature on race day and the winning race time depend on the level of relative humidity the day of the race? (Provide PRE, F\*, and a substantive interpretation.)
8. Assuming that the linear relationship between temperature and winning race time depends on relative humidity, what is the impact of temperature on winning times when relative humidity is 50%?
9. In model 6 where TIME is regressed on TEM, HUM and TEMHUM, the parameter estimate for HUM is nearly significant (PRE = .193;  $F^*(1, 17) = 4.14$ ;  $p = .058$ ). Provide a substantive interpretation of the null hypothesis that is being tested by these statistics.
10. Based on Models 6 and 7, is the product variable redundant with its components? Give a one-sentence interpretation or justification of your answer.
11. Based on the results presented in Models 2 through 7, write a paragraph that summarizes the apparent effects of the weather variables (race day high temperature and relative humidity) on winning race times.

Variable	N	Mean	Std Dev	Minimum	Maximum
TIME	21	130.3047619	1.7568370	128.0000000	134.9000000
TEM	21	63.0476190	9.8003887	49.0000000	80.0000000
HUM	21	0.5333333	0.1653280	0.3000000	0.9000000

Model: MODEL1  
 Dependent Variable: TIME

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
--------	----	----------------	-------------	---------	--------

Model	1	3.85747	3.85747	1.266	0.2744
Error	19	57.87206	3.04590		
C Total	20	61.72952			
Root MSE		1.74525	R-square	0.0625	
Dep Mean		130.30476	Adj R-sq	0.0131	
C.V.		1.33936			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
INTERCEP	1	271.013853	125.03473802	2.168	0.0431
YEAR	1	-0.070779	0.06289445	-1.125	0.2744

Model: MODEL2  
 Dependent Variable: TIME

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	31.55315	31.55315	19.867	0.0003
Error	19	30.17638	1.58823		
C Total	20	61.72952			
Root MSE	1.26025	R-square	0.5112		
Dep Mean	130.30476	Adj R-sq	0.4854		
C.V.	0.96716				

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
INTERCEP	1	122.224383	1.83361247	66.658	0.0001
TEM	1	0.128163	0.02875401	4.457	0.0003

Model: MODEL3  
 Dependent Variable: TIME

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	37.68364	18.84182	14.104	0.0002
Error	18	24.04589	1.33588		
C Total	20	61.72952			
Root MSE	1.15580	R-square	0.6105		
Dep Mean	130.30476	Adj R-sq	0.5672		
C.V.	0.88700				

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
INTERCEP	1	149.495282	12.84081376	11.642	0.0001
TEM	1	-0.743445	0.40772553	-1.823	0.0849
TEM2	1	0.006807	0.00317771	2.142	0.0461

Variable	DF	Type II SS	Squared Partial Corr Type II	Tolerance
INTERCEP	1	181.066516	.	.
TEM	1	4.441501	0.15591113	0.00418327
TEM2	1	6.130490	0.20315529	0.00418327

Model: MODEL4  
 Dependent Variable: TIME

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	37.68364	18.84182	14.104	0.0002
Error	18	24.04589	1.33588		
C Total	20	61.72952			
Root MSE	1.15580	R-square	0.6105		
Dep Mean	130.30476	Adj R-sq	0.5672		
C.V.	0.88700				

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
INTERCEP	1	129.682065	0.38484645	336.971	0.0001
TEMD	1	0.114929	0.02708495	4.243	0.0005
TEMD2	1	0.006807	0.00317771	2.142	0.0461

Variable	DF	Type II SS	Squared Partial Corr Type II	Tolerance
INTERCEP	1	151689	.	.
TEMD	1	24.052910	0.50007302	0.94797185
TEMD2	1	6.130490	0.20315529	0.94797185

Model: MODEL5  
 Dependent Variable: TIME

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	33.69546	16.84773	10.818	0.0008
Error	18	28.03406	1.55745		
C Total	20	61.72952			
Root MSE	1.24798	R-square	0.5459		
Dep Mean	130.30476	Adj R-sq	0.4954		
C.V.	0.95774				

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
INTERCEP	1	121.457648	1.92985971	62.936	0.0001
TEM	1	0.123410	0.02876102	4.291	0.0004
HUM	1	1.999566	1.70490905	1.173	0.2562

Variable	DF	Type II SS	Squared Partial Corr Type II	Tolerance
INTERCEP	1	6168.957728	.	.
TEM	1	28.674953	0.50565068	0.98014091
HUM	1	2.142312	0.07099303	0.98014091

Model: MODEL6  
 Dependent Variable: TIME

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	40.30655	13.43552	10.662	0.0004
Error	17	21.42297	1.26017		
C Total	20	61.72952			
Root MSE	1.12258	R-square	0.6530		
Dep Mean	130.30476	Adj R-sq	0.5917		
C.V.	0.86150				

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
INTERCEP	1	132.927945	5.30021629	25.080	0.0001
TEM	1	-0.068748	0.08779331	-0.783	0.4444
HUM	1	-18.234920	8.96640032	-2.034	0.0579
TEMHUM	1	0.337929	0.14753819	2.290	0.0350

Squared Partial Correlation and Tolerance					
Variable	DF	Type II SS	Corr Type II	Tolerance	
INTERCEP	1	792.640344	.	.	
TEM	1	0.772725	0.03481417	0.08511211	
HUM	1	5.211974	0.19568180	0.02867292	
TEMHUM	1	6.611092	0.23582355	0.01962461	

Model: MODEL7  
 Dependent Variable: TIME

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	40.30655	13.43552	10.662	0.0004
Error	17	21.42297	1.26017		
C Total	20	61.72952			
Root MSE	1.12258	R-square	0.6530		
Dep Mean	130.30476	Adj R-sq	0.5917		
C.V.	0.86150				

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
INTERCEP	1	130.231171	0.24706008	527.123	0.0001
TEMD	1	0.111470	0.02639092	4.224	0.0006
HUMD	1	3.070703	1.60331023	1.915	0.0724
TEMDHUMD	1	0.337929	0.14753819	2.290	0.0350

Squared Partial Correlation and Tolerance					
Variable	DF	Type II SS	Corr Type II	Tolerance	
INTERCEP	1	350151	.	.	
TEMD	1	22.482094	0.51206151	0.94190056	
HUMD	1	4.622441	0.17747620	0.89675345	
TEMDHUMD	1	6.611092	0.23582355	0.89395405	

### Question B

As part of his political platform, a candidate for political office in the South -- let's call him "Candidate A" -- claims that Southern schools are lagging behind other regions of the U.S. in educational spending and that this is the reason Southern students show poorer performance on standardized tests than do Northern students. He promises that if elected he will push education reform, and as a start he will increase the amount of money spent for education in his state. Candidate B argues, however, that the South spends just as much as the North does for education and thus that increased spending for education is unwarranted.

Your job is to determine if Candidate A's claim that southern states spend less on education than the northern states is justified. A commonly-used indicator of educational spending is the amount of money states spend on education per student each year (in thousands of dollars). Below you will find the average amount spent on education by states in the southeast, southwest, northeast, and northwest {note: only data from the 48 contiguous states are included in this analysis}.

	Northeast	Northwest	Southeast	Southwest
Mean:	4.21	3.63	2.97	3.24
StDev:	0.81	0.75	0.42	0.58
n:	15	12	12	9

1. Develop a full set of orthogonal contrast codes to analyze these data. Remember that one of the codes should address the claim made by Candidate A.
2. Calculate the parameter estimate for the contrast of amount spent by southern versus northern states. Provide a substantive interpretation of this parameter estimate.
3. Do southern and northern states spend different amounts on average for children's education? Be sure to include SSR, SSE(A), SSE(C), PRE and F\* in support of your conclusion.

### Question C

This question includes aspects of social geography, marketing, and social psychology. At issue are the number of trips made to shopping centers or malls. Each respondent keeps a diary for a year of all trips made to many different shopping centers. [There are nonindependence problems but we will ignore those until next semester. Just assume we are analyzing all the data for all the

respondents and all the shopping centers in one analysis.] The following variables are available for each respondent for each of a number of shopping centers.

TRIPS            number of trips by a respondent to a particular shopping center  
                    in the past year

DISTANCE      distance in miles from the residence of a respondent to a particular shopping center

MAJORS        number of major department stores in the center

SQFT            square feet of retail space

IN/OUT         whether the mall is enclosed (IN) or whether one goes outside  
                    between stores

CLEAN         respondent's rating of the mall's cleanliness

Specify the MODEL A/C combinations that you would use to answer each of the following questions . If you create any new variables, be sure to define them.

1.      Is the average number of trips by each respondent to each shopping center equal to 10?
2.      Does distance to a shopping center predict the respondent's number of trips to that shopping center?
3.      Assuming that distance is a useful predictor of number of trips, use a more powerful test to ask whether the average number of trips to each shopping center equals 10 (Question 1).
4.      As a group, do distance and the two size indicators—number of major department stores and number of square feet of retail space—predict number of trips?
5.      Over and above distance, do the two size measures predict number of trips?
6.      Just considering the square feet variable as the measure of size, does distance make less of a difference in trips to larger shopping centers than smaller shopping centers?
7.      In the context of the previous question, is there a significant effect of distance on the number of trips for malls having 500,000 square feet of retail space?

8. Just considering the square feet variable as the measure of size, is size a better predictor of trips for smaller shopping centers than for larger ones?
9. Are there more trips to enclosed malls than to outdoor shopping centers?
10. Controlling for distance to the mall, are there more trips to enclosed malls than to than outdoor shopping centers?
11. Does cleanliness make more of a difference for indoor malls than outdoor shopping centers?
12. Is the differential effect of distance for large and small shopping centers (Question 6) the same for indoor and outdoor malls? [Note: this is an extrapolation beyond any models we considered in class.]

### Question D

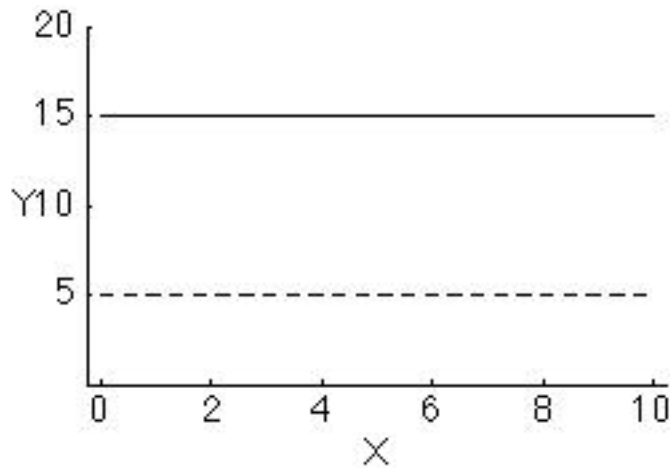
Each of the graphs below can be described by this equation:

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \beta_3 X_i Z_i + \beta_4 X_i^2 + e_i$$

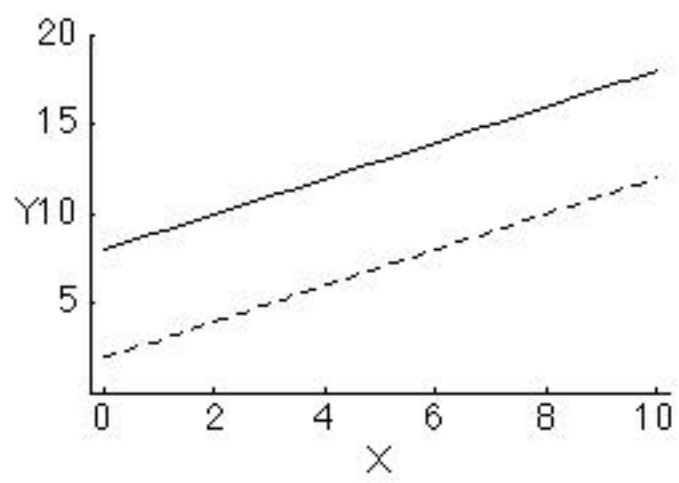
where  $Z_i$  is a coded predictor for a categorical variable such that +1 corresponds to the level indicated by the solid line in the graphs below and  $-1$  corresponds to the level indicated by the dashed line and where  $X_i$  is a continuous predictor with values between 0 and 10.

Each of the following graphs presents predicted values for data using the above equation. What we want you to do is to tell us about the values of the slope coefficients in the above equation for each of these graphs. Specifically, tell us for each slope coefficient (not the intercept) in the above model whether its value would be negative, zero, or positive. (Note: you do **not** need to specify the exact numerical value. Since there are four slopes, there are four parts to each of these questions. Accordingly each question, i.e., each graph, is worth four points.)

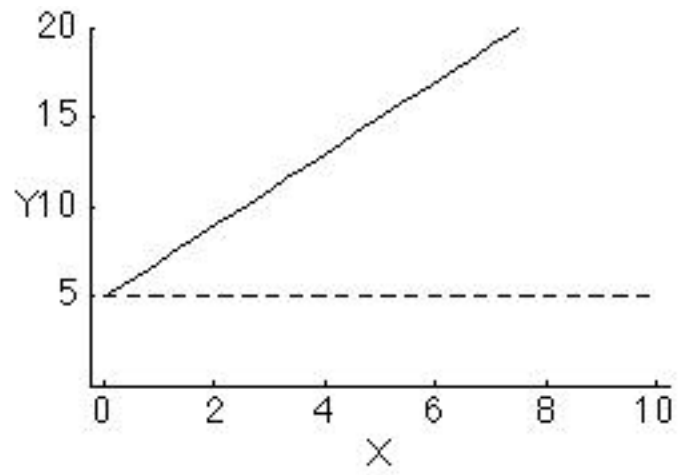
1.



2.



3.



4.

