

There are four questions on this exam. Except where noted, subquestions are worth 3 points each.

Question 1

A popular theory of sports performance is known as the "hot hand" hypothesis. It is most simply illustrated in the case of basketball. The hypothesis is that players are streak performers, that they can develop a "hot hand" and make nearly every shot in a row they take or they can develop a "cold hand" and miss a great number of shots in a row.

To evaluate the hot hand hypothesis, the sequential shooting records of 20 basketball players in the NBA were examined. Ten of these players were the individuals with the highest per game scoring average in the NBA. The other ten individuals were the ten players with the lowest per game scoring average. For each player, two probabilities were computed based on their sequential shooting record for the season:

PROB0: probability of making a shot given that the previous shot was a miss. (from 0 to 100)

PROB1: probability of making a shot given that the previous shot was made. (from 0 to 100)

The variable HILO represents whether the player is one of the 10 highest scorers (+1) or one of the ten lowest scorers (-1).

On the following pages are the means and proc reg outputs. The dependent variables in these analyses are defined as follows:

$$\begin{aligned} W0 &= (\text{PROB0} + \text{PROB1})/\text{sqrt}(2) \\ W1 &= (\text{PROB1} - \text{PROB0})/\text{sqrt}(2) \end{aligned}$$

Based on these analyses, answer the following questions:

1. Overall, across the 20 players, is there evidence to support the "hot hand" hypothesis? (Provide here only the F* and a substantive conclusion).
2. What are the values of SSR and PRE that correspond to the F* that you gave in your answer to question 1?
3. (6 points) Write an overall results section for these data, summarizing the analyses and interpreting the results. Present a graph of the cell means and discuss any and all significant differences.

4. You now want to test whether there is evidence for the "hot hand" hypothesis looking only at the data from the top 10 scorers (HILO=+1). What models C and A would you compare to answer this question? (Give parameter values.)

5. What is the value of the SSR for the model comparison given in response to question 4?

6. A transformation of these data may be in order prior to conducting the analysis. What problem would this transformation be likely to address and what transformation would you recommend?

7. Based on these data, it is tempting to conclude that overall (on average across ability levels of the players) the "hot hand" hypothesis is false. What might be wrong with this conclusion?

```
----- HILO=-1 -----
```

Variable	N	Mean	Std Dev	Minimum	Maximum
PROB0	10	30.7000000	2.8693786	26.0000000	35.0000000
PROB1	10	29.9000000	2.7264140	26.0000000	34.0000000

```
----- HILO=1 -----
```

Variable	N	Mean	Std Dev	Minimum	Maximum
PROB0	10	45.4000000	2.4585452	42.0000000	49.0000000
PROB1	10	47.4000000	3.5962944	40.0000000	53.0000000

Model: MODEL1
 Dependent Variable: W0

```
Analysis of Variance
```

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	2592.10000	2592.10000	191.220	0.0001
Error	18	244.00000	13.55556		
C Total	19	2836.10000			

Root MSE	3.68179	R-square	0.9140
Dep Mean	54.23509	Adj R-sq	0.9092
C.V.	6.78857		

```
Parameter Estimates
```

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	54.235090	0.82327260	65.877	0.0001
HILO	1	11.384419	0.82327260	13.828	0.0001

Model: MODEL2
Dependent Variable: W1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	19.60000	19.60000	5.204	0.0349
Error	18	67.80000	3.76667		
C Total	19	87.40000			
Root MSE	1.94079	R-square	0.2243		
Dep Mean	0.42426	Adj R-sq	0.1812		
C.V.	457.44864				

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	0.424264	0.43397389	0.978	0.3412
HILO	1	0.989949	0.43397389	2.281	0.0349

Question 2

Gary received this email message earlier this week:

> Sometimes extreme values of x induce extreme e 's (e.g., outliers). Right?

The e 's the writer is referring to are the errors or residuals for the model. Write a brief paragraph explaining to this person why they are wrong rather than "Right?" and what statistics could be used to separate these issues.

Question 3

[This problem context was suggested by a student in the class; it is loosely based on the work of Marty Seligman at Penn.] It is believed that depressed people tend to attribute good outcomes to luck or to the efforts of others (external locus of control) while they tend to view bad outcomes as consequences of their own behavior (internal locus of control). A researcher explores this idea by having a number of people complete two scales: a POSITIVE events scale and a NEGATIVE events scale. The items and scoring of the scales is such that high scores indicate attributions for the events to external causes while low scores indicate attributions to internal causes. The following additional variables are available for each participant in the study:

DEPRESS: score on the Beck Depression Inventory

SEX male or female

For each question, specify the appropriate Model A/C comparisons one would use to answer the question. Be sure to define any constructed variables you use.

1. Give the three Model A/C comparisons that would constitute a complete between-within ANOVA of the scale scores as a function of gender, whether the event is positive or negative, and their interaction.
2. Ignoring sex, does a higher score on the Beck Depression Inventory (BDI) predict a greater difference between the positive and negative scales?
3. Is the answer to the above question different for males than for females?
4. Ignoring sex again, the researcher wants to know if there is any difference between the positive and negative scale scores when scores on the BDI are at subclinical levels of about 18.
5. Still ignoring sex, does the difference between the positive and negative scales show a nonlinear relationship with depression score?
6. In the context of the previous model, the researcher wants to know if depression is still related to the difference when scores on the BDI are at subclinical levels of about 18.

There is causal ambiguity so it is just as reasonable to consider DEPRESS as a dependent variable instead of a predictor variable. For the remaining questions, use DEPRESS as the dependent variable.

7. Suppose the researcher randomly assigned participants to be in one of three treatment groups: a PLACEBO group which has weekly meetings with a therapist who gives no directive feedback, an ATTRIBUTIONAL therapy group in which training is given at weekly meetings to make attributions for bad outcomes to external causes, and a DRUG group that instead of weekly meetings receives standard medications for depression. Before actually doing the therapies, the researcher wants to check to make sure there are no pretreatment differences in the mean depression levels for any of the treatment groups or genders. Provide the Model A (you do NOT need to specify all the Model C's) you would use to perform a complete two-way between-subjects ANOVA to see if mean depression scores differed by treatment group and/or gender.
8. Ignoring treatment group and gender, do attributions for positive events predict depression over and above the attributions for negative events?
9. Still ignoring treatment group and gender, is depression predicted by the total and the difference of the two attributional scales?
10. Consider the values of R^2 that would be obtained for the Model A's in the previous two questions. We will refer to them as RSQ-8 and RSQ-9. Which of the following statements is true?

RSQ-8 > RSQ-9

RSQ-8 = RSQ-9

RSQ-8 < RSQ-9

Not enough information to tell

Briefly justify your answer.

Question 4

A school psychologist is interested in the relative concern students have of being popular in school. He administers a survey to girls and boys taken at random from elementary schools in each of the following settings: urban, suburban and rural. (Assume that data came from only one child from each school so that nonindependence is not a problem in these data.) The following variables are available in the dataset:

Variable	Description	Values
pgender	Participant gender	-1/2 for boys, +1/2 for girls
age	Participant age	Age in years
urbsub	Urban vs. Suburban	-1/2 for urban, +1/2 for suburban
ruroth	Rural vs. Other	+2/3 for rural, -1/3 for urban and suburban
goal	Student's personal goal	1 = goal is to be popular, 0 = don't care about being popular
genvsus	Pgender*Urbsub	
genvsro	Pgender*Ruroth	
genage	Pgender*age	

1. Overall, what proportion of students reported that their goal was to be popular?
2. Using the simplest model, did students from rural schools report different goals than students from urban and suburban schools? (Give χ^2 , p and a substantive conclusion.)
3. Using the simplest model, what is the probability that a girl in a suburban school reported that her goal was to be popular?
4. Allowing for gender by school type interactions, on average are girls more likely than boys to report that their goal is to be popular? (Give χ^2 , p and provide an interpretation of the relevant odds ratio.)

The LOGISTIC Procedure

Data Set: WORK.TEMP
 Response Variable: GOAL
 Response Levels: 2
 Number of Observations: 388
 Link Function: Logit

Response Profile

Ordered Value	GOAL	Count
1	1	141
2	0	247

The LOGISTIC Procedure

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	510.552	507.782	.
SC	514.513	519.665	.
-2 LOG L Score	508.552	501.782	6.770 with 2 DF (p=0.0339)
	.	.	6.896 with 2 DF (p=0.0318)

RSquare = 0.0173

Max-rescaled RSquare = 0.0237

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-0.5341	0.1070	24.9306	0.0001	.	.
URBSUB	1	0.0147	0.2558	0.0033	0.9543	0.003438	1.015
RUROTH	1	0.6045	0.2322	6.7807	0.0092	0.149147	1.830

The LOGISTIC Procedure

Data Set: WORK.TEMP
 Response Variable: GOAL
 Response Levels: 2
 Number of Observations: 388
 Link Function: Logit

Response Profile

Ordered Value	GOAL	Count
1	1	141
2	0	247

The LOGISTIC Procedure

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	510.552	505.110	.
SC	514.513	520.954	.
-2 LOG L Score	508.552	497.110	11.442 with 3 DF (p=0.0096)
	.	.	11.447 with 3 DF (p=0.0095)
	RSquare = 0.0291		Max-rescaled RSquare = 0.0398

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-0.5742	0.1099	27.3080	0.0001	.	.
URBSUB	1	0.1014	0.2605	0.1515	0.6971	0.023760	1.107
RUROTH	1	0.5712	0.2340	5.9611	0.0146	0.140924	1.770
PGENDER	1	0.4764	0.2220	4.6060	0.0319	0.130206	1.610

The LOGISTIC Procedure

Data Set: WORK.TEMP
 Response Variable: GOAL
 Response Levels: 2
 Number of Observations: 388
 Link Function: Logit

Response Profile

Ordered Value	GOAL	Count
1	1	141
2	0	247

The LOGISTIC Procedure

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	510.552	505.650	.
SC	514.513	529.416	.
-2 LOG L Score	508.552	493.650	14.902 with 5 DF (p=0.0108)
	.	.	14.306 with 5 DF (p=0.0138)

RSquare = 0.0377 Max-rescaled RSquare = 0.0516

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-0.5850	0.1130	26.7907	0.0001	.	.
URBSUB	1	0.2015	0.2736	0.5424	0.4615	0.047223	1.223
RUROTH	1	0.6902	0.2425	8.1037	0.0044	0.170288	1.994
PGENDER	1	0.4600	0.2260	4.1421	0.0418	0.125741	1.584
GENVSUS	1	-0.5423	0.5473	0.9818	0.3218	-0.062866	0.581
GENVSRO	1	-0.7642	0.4849	2.4834	0.1151	-0.094922	0.466

The LOGISTIC Procedure

Data Set: WORK.TEMP
 Response Variable: GOAL
 Response Levels: 2
 Number of Observations: 388
 Link Function: Logit

Response Profile

Ordered Value	GOAL	Count
1	1	141
2	0	247

The LOGISTIC Procedure

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	510.552	507.634	.
SC	514.513	535.361	.
-2 LOG L Score	508.552	493.634	14.918 with 6 DF (p=0.0209)
	.	.	14.322 with 6 DF (p=0.0262)

RSquare = 0.0377 Max-rescaled RSquare = 0.0516

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-0.7429	1.2476	0.3545	0.5516	.	.
AGE	1	0.0152	0.1193	0.0162	0.8989	0.007929	1.015
URBSUB	1	0.1918	0.2841	0.4561	0.4995	0.044952	1.211
RUROTH	1	0.6935	0.2438	8.0891	0.0045	0.171084	2.001
PGENDER	1	0.4603	0.2260	4.1461	0.0417	0.125809	1.585
GENVSUS	1	-0.5381	0.5483	0.9630	0.3264	-0.062375	0.584
GENVSRO	1	-0.7654	0.4850	2.4898	0.1146	-0.095063	0.465