Psychology 204b (Causal Modeling) Simonton Winter 2008

Exam II Your Name _____.

All questions in each part worth 10 points each.

Part I: Multiple Regression Models

Below are data from a hypothetical survey of science professors. The dependent variable is the number of citations they have received in professional journals in the past five years. The predictors are gender (1 = male and 2 = female), discipline (1 = physical sciences, 2 = biological sciences, and 3 = social sciences), age, and achievement motivation (*n*Ach).

Citations	Gender	Discipline	Age	*n*Ach
55	1	2	42	30
86	2	2	69	90
97	1	3	35	10
120	2	3	60	60
101	1	1	40	50
13	1	2	51	40
44	2	1	37	70
169	1	3	58	80
78	2	1	71	20
50	1	1	65	50
21	1	2	53	60
12	2	3	46	40

1) Assuming that you want to use the male physical scientists as the comparison group, please specify the data transformations necessary to put the variables in a form more appropriate for the regression analyses required below. You may use the syntax any standard statistical software that suits your fancy. Be sure to give all variables reasonable names for use in the computer analyses. (2 points for each numerical variable and 3 points for each categorical variable)

For each of the questions that follow: (a) show the regression equation that defines the hypothesis being addressed and (b) show specifically how the resulting regression coefficients are to be properly interpreted. Hence, all of the following questions have two components (5 points each):

2) How would you test the main effects of Gender and Discipline?

3) How would you test for Gender X Discipline interaction effects?

4) How would the main effects of Gender and Discipline be altered if we introduced Age and *n*Ach as controls or "covariates"? How do we determine whether the effects of the controls are the same across both Gender and Discipline?

5) How can we detect whether the impact of *n*Ach varies according to Age? In particular, what if *n*Ach has a positive predictive value for younger scientists (Age below the mean of Age) but a negative predictive value for older scientists (Age above the mean of Age)?

6) How can you determine whether the impact of Age was actually curvilinear? In particular, what if you suppose that those who are middle-aged (at the mean of Age) are likely to be the most productive?

7) Finally, what if you thought that the above curvilinear function varied according Gender? Specifically, what if you hypothesized that men peaked earlier than women?

Part II: Analytical Complications

Part II: Analytical Complications

This analysis predicts gross US box office as a function of production budget, number of screens on opening weekend, sequel, film critic ratings (Metacritic), best picture awards and nominations, and two relevant MPAA ratings.

1) Are there any problems here with autoregressive residuals? How do you know? (5 points) What are the consequences whenever such autoregression is substantial? (3 points) What are the likely consequences in the present case? (2 points)

2) Are there any problems here with heteroscedasticity? How do you know? (5 points) What are the consequences whenever heteroscedasticity is substantial? (3 points) Any recommendation about how to lessen this problem here? (2 points)

3) Are there any prominent outliers? What kinds of problems do outliers represent? (3 points) What do the leverage values tell us? (3 points) What is the fundamental difference between the measures of outliers and the measures of leverage? (2 points) How might these two measures be used together to make important inferences about the influence that certain scores have on the regression results? Are there any problems in the present analysis? (2 points)

Dep Var: GROSS N: 371 Multiple R: 0.7467 Squared multiple R: 0.5576

Adjusted squared multiple R: 0.5491 Standard error of estimate: 55.7197

Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	-110.1014	12.3676	0.0000	.	-8.9024	0.0000
BUDGET	0.6417	0.1016	0.3108	0.5033	6.3173	0.0000
SCREENS	0.0262	0.0033	0.4254	0.4233	7.9280	0.0000
SEQUEL	45.8301	12.8856	0.1306	0.9035	3.5567	0.0004
METACRIT	1.5509	0.1732	0.3690	0.7176	8.9540	0.0000
BEST	41.7371	15.2466	0.1079	0.7838	2.7375	0.0065
PG	31.0558	10.0438	0.1136	0.9023	3.0920	0.0021
PG13	14.4514	6.6994	0.0807	0.8717	2.1571	0.0317

Analysis of Variance

Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
Regression	1.42060E+06	7	202942.6552	65.3665	0.0000
Residual	1.12700E+06	363	3104.6904		

*** WARNING ***

Case 3781 is an outlier (Studentized Residual = 4.4250)
Case 3843 has large leverage (Leverage = 0.1199)
Case 4358 is an outlier (Studentized Residual = 4.0657)
Case 4378 is an outlier (Studentized Residual = 4.0658)
Case 4618 has large leverage (Leverage = 0.1293)
Case 4849 is an outlier (Studentized Residual = 5.1297)
Case 4858 is an outlier (Studentized Residual = 8.3251)

Durbin-Watson D Statistic 2.0592
First Order Autocorrelation -0.0307

Plot of residuals against predicted values