Department of Mathematics and Statistics

The University of Toledo

Master's Comprehensive Examination
Applied Statistics

November 23, 2013

Instructions:

Do all four problems;

Show all of your computations;

Prove all of your assertions or quote appropriate theorems;

This is three-hour open book examination.

1. (20 points) In an early study of the effects of a strong magnetic field on the development of mice, 7 cages, each containing 3 albino female mice were subjected for a period of ten days to a magnetic field. 21 other mice housed in 7 similar cages were not placed in a magnetic field and served as controls. The following table shows the weight gains, in grams, for each of the cages.

Magnetic Field Present: 22.8 10.2 20.8 27.0 19.2 10.4 14.2

Magnetic Field Absent: 23.6 31.0 19.5 26.2 26.5 25.2 24.5

Questions:

a. (10 points) State a nonparametric model, define hypotheses, and carry out a test at level $\alpha = 0.10$ which will enable you to decide whether there is a significant difference in weight gain between these two groups. Find the exact p-value and use it to make your decision.

b. (10 points) Repeat part a using the normal approximation. Do you get the same conclusion from (a)?

2. (30 points) The data have id-numerical identifier for each student; dependent variable: GPA-dependent variable, the grade point average after three semesters; six explanatory variables: HSM, HSS, HSE, SATM, SATV and GENDER - coded as 1 for men and 2 for women. The following is part of the data collected:

ID	GPA	HSM	HSS	HSE	SATM	SATV	GENDER
001	3.32	10	10	10	670	600	1
002	2.26	6	8	5	700	640	1
003	2.35	8	6	8	640	530	1
004	2.08	9	10	7	670	600	1
005	3.38	8	9	8	540	580	1
006	3.29	10	8	8	760	630	1
007	3.21	8	8	7	600	400	1
008	2.00	3	7	6	460	530	1
009	3.18	9	10	8	670	450	1
010	2.34	7	7	6	570	480	1
...							
223	2.59	5	4	7	630	470	2
224	2.28	9	8	9	559	488	2

Based on the SAS output in the appendix , answer the following questions: Questions:

a. (5 points) Give the equation of the fitted regression line using all six explanatory variables.

In the following problems, use the model which only have HSM and HSE as explanatory variables to predict the response GPA.

b. (5 points) Obtain the variance inflation factors. Are there indications that serious multicollinearity problems exist here? Explain.

c. (5 points) Obtain the studentized deleted residuals for observation 8. Use the Bonferroni outlier test procedure with $\alpha = 0.10$ to identify whether it is an outlying Y oberservation.

d. (5 points) Use the diagonal elements of the hat matrix to identify whether observation 8 is an outlying X observation. State the decision rule and conclusion.

e. (10 points) Obtain DFFITS, DFBETAS and Cook's distance values for observation 8 to assess its influence. What do you conclude? (Hint: $F_{3,221}(0.5) = 0.7911$)

3. (25 points) The research interest is the development of bronchopulmonary dysplasia (BPD). A sample of 223 infants participated in the study. The response variable `bpd` is binary, denoting whether an infant develops BPD (`bpd`=0 if no; `bpd`=1 if yes) by day 28 of life. The covariates are `weight` (birth weight in grams), `gestage` (gestational age in weeks), `toxemia` (`toxemia` =1 if toxemia is present; `toxemia`=0 if toxemia is absent). The following is part of the data collected:

bpd	weight	gestage	toxemia
1	8.5	27	0
0	15.0	33	0
1	13.6	32	0
0	9.6	35	1
0	15.6	33	0
0	11.2	29	0
1	8.1	28	0
0	16.2	32	0
1	10.0	30	0
1	7.0	26	0
⋮			

Questions:

a. (10 points) What is the appropriate full model for this study? Use dummy coding for discrete covariates. Label the variables in the model clearly.

Based on the SAS output in the appendix for the full model which includes the three covariates, answer the following two questions:

b. (5 points) Can the full model be reduced? *Briefly explain.*

c. (10 points) Suppose the output is obtained based on your full model in (a). How do you interpret "−0.2644" and "−1.3438" in the output.

4. (25 points) In a clinical trial of patients suffering from epileptic seizures, patients were randomized to receive either a placebo or the drug in addition to the therapy. The goal of the analysis is to compare the two groups, in particular to answer the question whether the drug is effective in reducing epileptic seizures. The response variable `y` is the number of epileptic seizure by the end of 8-week treatment period. The covariates are `trt` (`trt`=0 for the placebo group; `trt`=1 for the drug group) and `age` (age of the patient). The following is part of the data collected:

trt	age	y
0	31	3
0	30	3
0	25	0
0	36	1
0	22	9
1	18	9
1	32	9
1	20	3
1	20	1
1	18	7
⋮		

Questions:

a. (10 points) What is the appropriate full model for this study? Use dummy coding for discrete covariates. Label the variables in the model clearly.

Based on the SAS output in the appendix for the models with dispersion and without dispersion, answer the following questions:

b. (5 points) Do you think that the dispersion should be accounted for? *Briefly explain.*

c. (10 points) Using the model that accounts for the dispersion, what is your conclusion about the effect of the drug in reducing the number of epileptic seizures.

Appendix for Problem 2

SAS Output

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: GPA

Number of Observations Read 224
Number of Observations Used 224

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	28.68514	4.78086	9.72	<.0001
Error	217	106.77765	0.49206		
Corrected Total	223	135.46279			

Root MSE	0.70147	R-Square	0.2118
Dependent Mean	2.63522	Adj R-Sq	0.1900
Coeff Var	26.61907		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.27864	0.43369	0.64	0.5212
HSM	1	0.14423	0.03979	3.62	0.0004
HSS	1	0.03827	0.03874	0.99	0.3244
HSE	1	0.05103	0.04228	1.21	0.2287
SATM	1	0.00100	0.00071725	1.40	0.1633
SATV	1	-0.00041086	0.00059323	-0.69	0.4893
GENDER	1	0.03237	0.11148	0.29	0.7718

The REG Procedure
Model: MODEL1
Dependent Variable: GPA

Number of Observations Read 224
Number of Observations Used 224

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	27.30349	13.65175	27.89	<.0001
Error	221	108.15930	0.48941		
Corrected Total	223	135.46279			

Root MSE	0.69958	R-Square	0.2016
Dependent Mean	2.63522	Adj R-Sq	0.1943
Coeff Var	26.54718		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	0.62423	0.29172	2.14	0.0335	.	0
HSM	1	0.18265	0.03196	5.72	<.0001	0.80029	1.24954
HSE	1	0.06067	0.03473	1.75	0.0820	0.80029	1.24954

Model: MODEL1
Dependent Variable: GPA

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	3.3200	3.0575	0.0792	0.2625	0.695	0.378	| | |	0.001
2	2.2600	2.0235	0.1100	0.2365	0.691	0.342	| | |	0.001
3	2.3500	2.5708	0.0477	-0.2208	0.698	-0.316	| | |	0.000
4	2.0800	2.6928	0.0695	-0.6128	0.696	-0.880	| * | |	0.003
5	3.3800	2.5708	0.0477	0.8092	0.698	1.159	| |** |	0.002
6	3.2900	2.9361	0.0723	0.3539	0.696	0.509	| |* |	0.001
7	3.2100	2.5102	0.0582	0.6998	0.697	1.004	| |** |	0.002
8	2.0000	1.5362	0.1592	0.4638	0.681	0.681	| |* |	0.008

Dependent Variable: GPA

Output Statistics

Obs	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	DFBETAS HSM	DFBETAS HSE
1	0.3770	0.0128	1.0249	0.0429	-0.0299	0.0130	0.0229
2	0.3416	0.0247	1.0378	0.0544	0.0509	-0.0129	-0.0367
3	-0.3157	0.0046	1.0170	-0.0216	-0.0064	0.0040	-0.0006
4	-0.8799	0.0099	1.0131	-0.0879	-0.0230	-0.0489	0.0603
5	1.1603	0.0046	1.0000	0.0792	0.0237	-0.0147	0.0022
6	0.5077	0.0107	1.0211	0.0528	-0.0120	0.0402	-0.0199
7	1.0039	0.0069	1.0069	0.0838	0.0483	0.0097	-0.0481
8	0.6800	0.0518	1.0623	0.1589	0.1295	-0.1373	0.0033

Appendix for Problem 3

SAS Output

The GENMOD Procedure

Model Information

Data Set	WORK.LOWBWT
Distribution	Binomial
Link Function	Logit
Dependent Variable	bpd
Number of Observations Read	223
Number of Observations Used	223
Number of Events	76
Number of Trials	223

Response Profile

Ordered Value	bpd	Total Frequency
1	1	76
2	0	147

PROC GENMOD is modeling the probability that bpd='1'.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	13.9361	2.9826	8.0904	19.7818	21.83	<.0001
weight	1	-0.2644	0.0812	-0.4236	-0.1051	10.59	0.0011
gestage	1	-0.3885	0.1149	-0.6137	-0.1634	11.44	0.0007
toxemia	1	-1.3438	0.6075	-2.5345	-0.1531	4.89	0.0270
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Appendix for Problem 4

SAS Output Without Dispersion

The GENMOD Procedure

Model Information

Data Set	WORK.SEIZURE
Distribution	Poisson
Link Function	Log
Dependent Variable	y
Number of Observations Read	59
Number of Observations Used	59

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	56	783.7476	13.9955
Scaled Deviance	56	783.7476	13.9955
Pearson Chi-Square	56	1346.9467	24.0526
Scaled Pearson X2	56	1346.9467	24.0526
Log Likelihood		567.2587	
Full Log Likelihood		-482.7089	
AIC (smaller is better)		971.4177	
AICC (smaller is better)		971.8541	
BIC (smaller is better)		977.6503	

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	2.5906	0.1951	2.2083	2.9729	176.41	<.0001
trt	1	-0.0861	0.0898	-0.2621	0.0899	0.92	0.3378
age	1	-0.0145	0.0065	-0.0272	-0.0019	5.06	0.0246
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

SAS Output With Dispersion

The GENMOD Procedure

Model Information

Data Set	WORK.SEIZURE
Distribution	Poisson
Link Function	Log
Dependent Variable	y
Number of Observations Read	59
Number of Observations Used	59

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	56	783.7476	13.9955
Scaled Deviance	56	783.7476	13.9955
Pearson Chi-Square	56	1346.9467	24.0526
Scaled Pearson X2	56	1346.9467	24.0526
Log Likelihood		567.2587	
Full Log Likelihood		-482.7089	
AIC (smaller is better)		971.4177	
AICC (smaller is better)		971.8541	
BIC (smaller is better)		977.6503	

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	2.5906	0.7297	1.1605	4.0208	12.60	0.0004
trt	1	-0.0861	0.3360	-0.7446	0.5724	0.07	0.7978
age	1	-0.0145	0.0242	-0.0619	0.0328	0.36	0.5478
Scale	0	3.7411	0.0000	3.7411	3.7411		

NOTE: The scale parameter was estimated by the square root of DEVIANCE/DOF.