MASTER'S DIAGNOSTIC EXAMINATION - August 20, 2014

Student's Name ____

INSTRUCTIONS FOR STUDENTS:

1. The exam is to be started at Noon (CDT) and completed by 4 pm (CDT) on August 20, 2014.
2. Put your name above but DO NOT put your NAME on the **SOLUTIONS** to the exam.
3. Place the NUMBER assigned to you on the

   UPPER RIGHT HAND CORNER of EACH PAGE of your SOLUTIONS.
4. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
5. Use only one side of each sheet of paper.
6. You must answer all four questions: Questions I, II, III and IV.
7. Be sure to attempt all parts of the four questions. It may be possible to answer a later part of a question without having solved the earlier parts.
8. Be sure to hand in all of your exam. No additional material will be accepted once the exam has ended and you have left the exam room.
9. You may use the following:
   - Calculator which does not have capability to phone, text, or access the Web
   - Pencil or pen
   - Blank paper for the solutions for this examination
   - No other materials are allowed

- I attest that I spent no more than 4 hours to complete the exam.
- I used only the materials described above.
- I did not receive assistance from anyone during the taking of this exam.

Student's Signature ____________________________________

INSTRUCTIONS FOR PROCTOR:

Immediately after the student completes the exam, **fax** the student's solutions to **979-845-6060** or

Scan the solutions into a **single** pdf file and **email to longneck@stat.tamu.edu**

Do not send the questions, just send the student's solutions.

(1) I certify that the time at which the student started the exam was ____________________

and the time at which the student completed the exam was ____________________

(2) I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.

(3) I certify that the student's solutions were faxed to **979-845-6060** or

emailed to **longneck@stat.tamu.edu**.

Proctor's Signature____________________________________

QUESTION I. Part A:

For the following experiment, provide the following information:

1. Type of Randomization, for example, CRD, RCBD, LSD, BIBD, SPLIT-PLOT, Crossover, etc.;
2. Type of Treatment Structure, for example, Single Factor, Crossed, Nested, Fractional, etc.;
3. Identify each of the factors as being Fixed or Random;
4. Describe the Experimental Units and Measurement Units;
5. Describe the Measurement Process: Response Variable, Covariates, SubSampling, Repeated Measures;
6. A partial ANOVA Table containing Sources of Variation (SV), Degrees of Freedom, and Expected Mean Squares;
7. Write a statistical model for this experiment and include all necessary conditions on the model parameters and variables.

A leading brand of ice cream designs an experiment to evaluate the impact of several artificial sweeteners on the texture of their product. It is well known that replacing natural sweeteners with artificial sweeteners in ice cream can result in a product which is has an unappealing texture. A proposed method to overcome this problem is to increasing the blending time in the production process. The researchers decided to use four types of sweeteners: a natural sweetener (Control), Aspartame, Saccharin, and Sucralose. Twelve containers of ice cream were made, three of each of the four types of sweeteners, with the type of sweetener randomly assigned to the containers. Each of the 12 containers of ice cream was then split into four portions. The four portions are then randomly assigned to one of four blending times: 1 minute, 2 minutes, 5 minutes, and 8 minutes. At the end of the specified blending period, the ice cream is assigned a texture score. The researcher was particularly interested in the impact of the four sweeteners and the blending times on the average texture scores.

		Blending Time(min.)			
Sweetener	Container	1	2	5	8
Control	1	7	10	17	22
	2	4	4	11	23
	3	4	11	10	31
Aspartame	1	8	12	22	27
	2	6	7	27	30
	3	9	8	29	32
Saccharin	1	7	8	21	35
	2	1	4	13	25
	3	5	4	13	28
Sucralose	1	3	11	21	37
	2	1	12	25	31
	3	4	9	27	32

PROBLEM I. Part B:

For each of the following questions, select **ONE** letter from the list on the next page which is the **BEST** solution to each of the following situations.

SITUATION:

(1) A CRD with three factors: F_1-fixed levels, F_2-fixed levels, F_3-fixed levels, was conducted. The experimenter obtained the following results from the AOV F-tests: $F_1 * F_2 * F_3$ is significant, $F_1 * F_2$-not significant, $F_1 * F_3$-significant, $F_2 * F_3$- not significant, and F_1, F_2, F_3 are all not significant. She wants to determine which pairs of means are different across the levels of F_1.

(2) A three factor experiment is run with Factor F_1-fixed, Factor F_2-fixed having levels nested within the levels of Factor F_1, Factor F_3-fixed crossed with Factor F_1. The interaction between factors F_1 and F_3 was not significant and the interaction between factors $F_2(F_1)$ and F_3 was not significant. The researcher was interested in determining which pairs of means are different across the levels of F_1.

(3) A RCBD with three factors: F_1-fixed, F_2-fixed, F_3-random, was conducted. The experimenter obtained the following results from the AOV F-tests: $F_1 * F_2 * F_3$ is not significant, $F_1 * F_2$-not significant, $F_1 * F_3$-significant, $F_2 * F_3$-significant, and F_1, F_2, F_3 are all not significant. She wants to determine if there are pairwise differences in the levels of F_1.

(4) In a quality control experiment, the production engineer was interested in evaluating factors which may have caused a high defective rate in a product. There are five Rates, F_1, at which a platinum coating is applied to the product, with levels, 1.0, 1.1, 1.2, 1.3, 1.4 mm/second. The second factor, F_2, is four Types of Machines used to apply the coating to the product, with levels, M1, M2, M3, M4. The third factor, F_3, was the Operators the coating machines. Twenty operators of the coating machines were randomly selected from the workforce. Each operator applied the coating to 80 units of the product, four units for each combination of a Rate and Type of Machine. There was significant evidence of a 3-factor interaction and all 2-factor interactions were found to be significant. The company wants to know if the mean defective rate, μ_{ijk}, increased as the Rate, F_1, of applying the coating was increased.

(5) An experiment is designed to investigate plant growth involving a factorial treatment structure with factor F_1, the temperature in a growth chamber, $15°C$, $20°C$, $30°C$, $35°C$ crossed with factor F_2, four brands of growth stimulants at three dose levels: 0 ml/mg, 10 ml/mg, 15 ml/mg, factor F_3. The experiment was conducted as a completely randomized design with 10 flowers randomly assigned to each of the 36 treatments. The experimenter determined from the AOV F-tests that only the following effects were significant: $F_1 * F_3$, $F_1 * F_2$, F_2, F_3. The researcher wants to determine the temperature that yields the maximum mean growth.

TECHNIQUE:

A. Trend analysis using Scheffe contrasts

B. Trend analysis using Bonferroni contrasts

C. Trend analysis in the levels of F_1 averaged over levels of the other factors

D. Trend analysis in the levels of F_1 separately at each level of the other factors

E. Trend analysis in the levels of F_1 separately at each level of F_2 but averaged over the other factors

F. Scheffe's test for contrast differences

G. Dunnett's comparison technique

H. Dunnett's comparison technique to all combinations of the factors

I. Dunnett's comparison technique applied to the levels of factor F_1 separately at each level of the other factors

J. Dunnett's comparison technique applied to the levels of factor F_1 averaged over the levels of the other factors

K. Tukey's comparison technique

L. Tukey's comparison technique to all combinations of the factors

M. Tukey's comparison technique applied to the levels of factor F_1 separately at each level of the other factors

N. Tukey's comparison technique applied to the levels of factor F_1 averaged over the levels of the other factors

O. Hsu's comparison technique

P. Hsu's comparison technique applied to the levels of factor F_1 separately at each level of the other factors

Q. Hsu's comparison technique applied to the levels of factor F_1 averaged over the levels of the other factors

R. Hsu's comparison technique applied to all combinations of the factors

S. Nothing new is learned beyond the results of the F-tests from the AOV table.

T. Comparison of marginal means is not appropriate.

U. None of the above methods are appropriate.

QUESTION II.

Consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, 2, \ldots, n$$

Suppose we observe the following values for the y_i and x_i:

y	x
0.34	1
1.28	2
1.16	3
2.11	4
2.66	5

Recall that $(\hat{\beta}_0, \hat{\beta}_1)' = \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, where $\boldsymbol{X}$ is the model matrix and that $\mathrm{Var}(\boldsymbol{v}'\boldsymbol{Y}) = \boldsymbol{v}'\mathrm{Var}(\boldsymbol{Y})\boldsymbol{v}$, where $\boldsymbol{v}$ is a vector of constants and $\boldsymbol{Y}$ is a random vector.

1. Compute $\hat{\boldsymbol{\beta}}$.

2. Compute $\hat{\sigma}$.

3. Compute the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$.

4. Compute a 95% confidence interval for the mean response when $x = 3$.

5. Compute a 95% prediction interval for a new observation when $x = 3$.

QUESTION III.

Part A. The safety of people who live or work near nuclear-power plants has been under debate in recent years. One possible health hazard is an excess number of deaths due to cancer among those exposed. A problem with studying this is that the number of deaths from cancer is small, making it difficult to make statistical conclusions. An alternative approach used by epidemiologists is the *proportional mortality study*, in which the proportion of deaths in the study group is compared with the corresponding proportion in a large population. Suppose that 15 deaths have occurred among 55- to 64-year-old male workers in a nuclear power plant during a given time period and that in 6 of them the cause of death was cancer. In the general population of 55- to 64-year-old males it is known that approximately 20% of all deaths can be attributed to cancer.

1. Formulate the hypotheses, compute a $p-$value, and state a conclusion at the 5% level for a test of whether the true proportion of 55- to 64-year-old male workers in a nuclear power plant having deaths attributable to cancer exceeds 20%. You may find it useful to use the attached tables of the binomial distribution.

2. Statisticians have found that using one-sided tests based upon a test statistic with a discrete distribution can be overly conservative (i.e., losing power by failing to reject too often for a test of the desired level of significance). To help adjust for this phenomenon, one can use the mid $P-$value for a one-sided test. The mid $P-$value equals one-half the probability of the observed result plus the probability of the more extreme results. Compute a the mid $P-$value, and state a conclusion at the 5% level for a test of whether the true proportion of 55- to 64-year-old male workers in a nuclear power plant having deaths attributable to cancer exceeds 20%.

3. Taking into account the sample size, construct a 95% confidence interval for the true proportion of 55- to 64-year-old male workers in a nuclear power plant having deaths attributable to cancer.

Part B. The epidemiologist also considered data for all 55- to 64-year-old male workers in the local area. Suppose that 254 deaths occurred among 55- to 64-year-old male workers during the given time period and that in 59 of them the cause of death was cancer.

1. Formulate the hypotheses, compute a $p-$value, and state a conclusion at the 5% level for a test of whether the true proportion of 55- to 64-year-old male workers in the local area having deaths attributable to cancer exceeds 20%.

2. Taking into account the sample size, construct a 95% confidence interval for the true proportion of 55- to 64-year-old male workers in the local area having deaths attributable to cancer.

3. Give a brief explanation why you were able to answer the two previous parts of this problem without using binomial tables.

QUESTION IV.

Chapter 2 of Miller (2013) Modelling Techniques in Predictive Analytics, Pearson, New Jersey makes extensive use of multiple regression to analyze attendance figures for the 81 Dodgers home games in the 2012 Major League Baseball season. Data are available on the following variables

- Attendance = home game attendance (i.e., the number of tickets sold to each game)
- Month = month in which each game was played
- Day_of_week = day of the week each game was played on
- OpponentsFromLargeMetroAreas = a dummy variable which is 1 if the opponent is the New York Mets, Chicago Cubs and White Sox, Los Angeles Angels and the Washington DC Nationals
- Temp = temperature at the stadium during the game
- Day_night = day (for day games) and night (for night games)
- BobbleheadPromotion = a dummvariable which is 1 if the game involved a bobblehead promotion

According to Miller (2013):

"Dodger Stadium, with a capacity of 56,000, is the largest ballpark in the world. From the data, we can see that Dodger Stadium was filled to capacity only twice in 2012. . . . The eleven bobblehead promotions occurred on night games, six of those being Tuesday nights. . . . Opponents from the large metropolitan areas (the New York Mets, Chicago Cubs and White Sox, Los Angeles Angels and Washington D.C. Nationals) are consistently associated with higher attendance. . . . Explanatory graphics help us find models that might work for predicting attendance and for evaluating the effect of promotions on attendance.

Figure 2.1 shows distributions of attendance across days of the week, and

Figure 2.2 shows attendance by month.

To advise management regarding (bobblehead) promotions, we would like to know if promotions have a positive effect upon attendance, and if they do have a positive effect, how much that effect might be. To provide this advice we build a linear model for predicting attendance using month, day of the week and an indicator variable for the bobblehead promotion . . . "

Given on the next few pages are Figures 2.1 and 2.2, JMP output from the least squares model fit by Miller (2013) and some additional plots. A statistics professor originally from Australia has taken a careful look at the data and found among other things that there is no evidence of significant autocorrelation in the attendance results. In other words, there is no evidence that attendance at home games on day t is statistically significantly related to attendance on days $t-1,\ t-2,\ \ldots$.

1. Describe in detail one major concern that potentially threatens the validity of the model fit by Miller (2013).
2. Explain the specific steps you would take to overcome the problem described in part (1).
3. On the basis of the plots presented what predictors would you recommend being included in the model you describe in part (2).

Figure 2.1

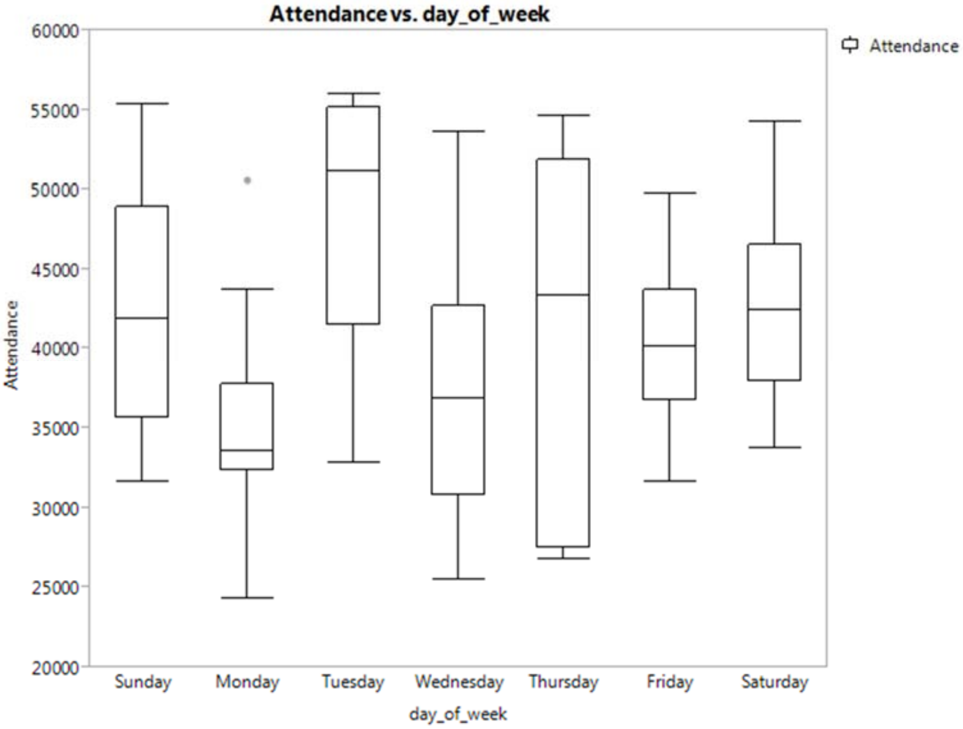
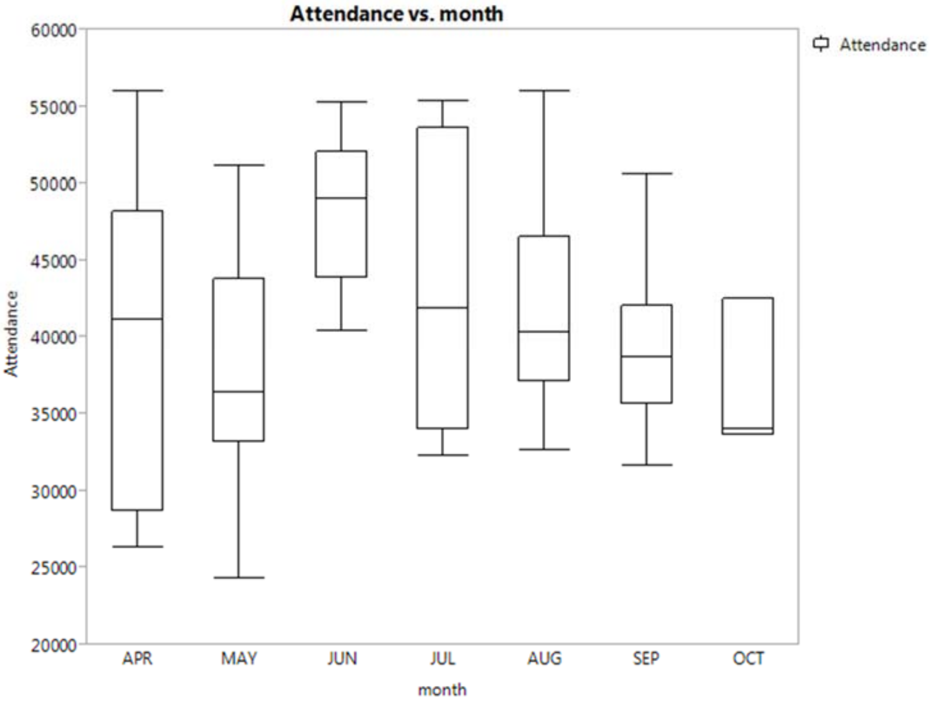
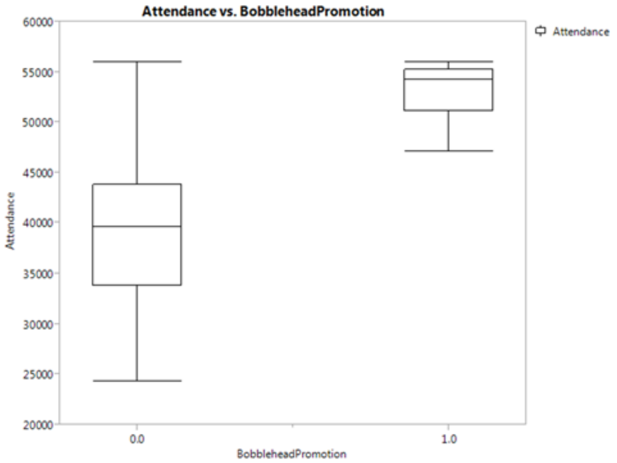
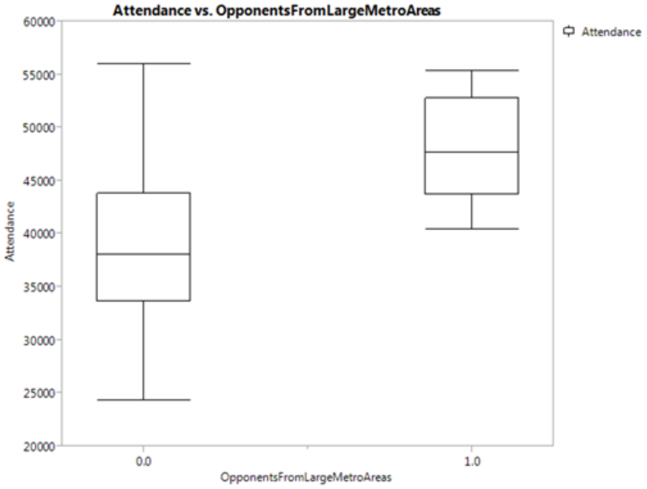


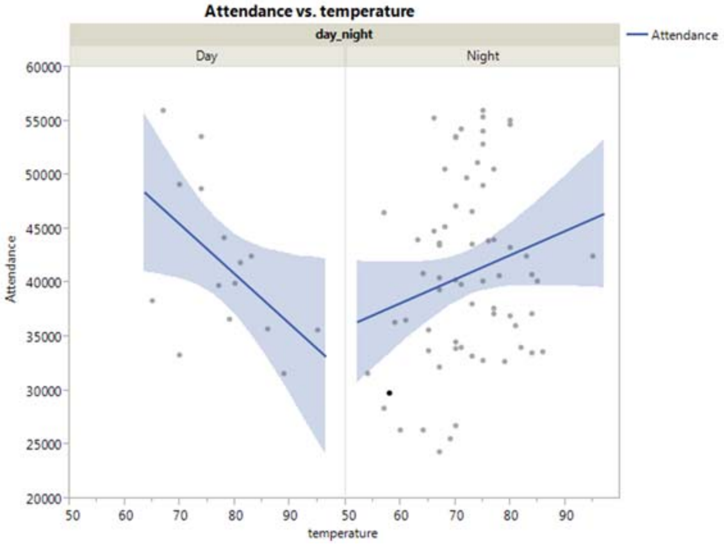
Figure 2.2



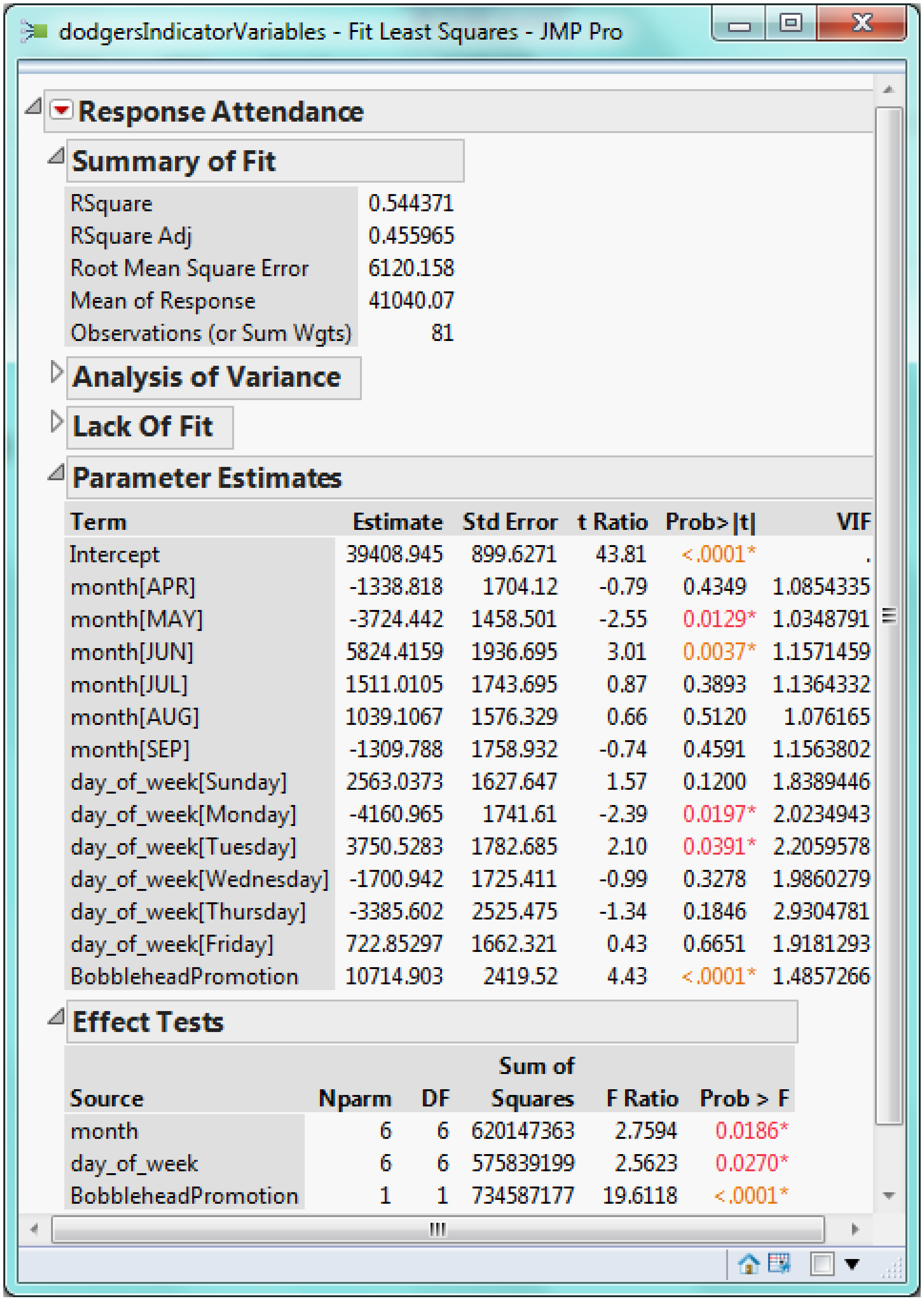
Other plots not in Miller (2013)







Model fit by Miller (2013)



dodgersIndicatorVariables - Fit Least Squares - JMP Pro

Response Attendance

Summary of Fit

RSquare	0.544371
RSquare Adj	0.455965
Root Mean Square Error	6120.158
Mean of Response	41040.07
Observations (or Sum Wgts)	81

Analysis of Variance

Lack Of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	39408.945	899.6271	43.81	<.0001*	.
month[APR]	-1338.818	1704.12	-0.79	0.4349	1.0854335
month[MAY]	-3724.442	1458.501	-2.55	0.0129*	1.0348791
month[JUN]	5824.4159	1936.695	3.01	0.0037*	1.1571459
month[JUL]	1511.0105	1743.695	0.87	0.3893	1.1364332
month[AUG]	1039.1067	1576.329	0.66	0.5120	1.076165
month[SEP]	-1309.788	1758.932	-0.74	0.4591	1.1563802
day_of_week[Sunday]	2563.0373	1627.647	1.57	0.1200	1.8389446
day_of_week[Monday]	-4160.965	1741.61	-2.39	0.0197*	2.0234943
day_of_week[Tuesday]	3750.5283	1782.685	2.10	0.0391*	2.2059578
day_of_week[Wednesday]	-1700.942	1725.411	-0.99	0.3278	1.9860279
day_of_week[Thursday]	-3385.602	2525.475	-1.34	0.1846	2.9304781
day_of_week[Friday]	722.85297	1662.321	0.43	0.6651	1.9181293
BobbleheadPromotion	10714.903	2419.52	4.43	<.0001*	1.4857266

Effect Tests

			Sum of		
Source	Nparm	DF	Squares	F Ratio	Prob > F
month	6	6	620147363	2.7594	0.0186*
day_of_week	6	6	575839199	2.5623	0.0270*
BobbleheadPromotion	1	1	734587177	19.6118	<.0001*

Table A.1 Cumulative Binomial Probabil

itie

s (*cont.*)

c. $n = 15$

	0.01	0.05	0.10	0.20	0.25	0.30
0	0.860	.463	.206	.035	.013	.00
1	.990	.829	.549	.167	.080	.03
2	1.000	.964	.816	.398	.236	.12
3	1.000	.995	.944	.648	.461	.29
4	1.000	.999	.987	.836	.686	.51
5	1.000	1.000	.998	.939	.852	.72
6	1.000	1.000	1.000	.982	.943	.86
x 7	1.000	1.000	1.000	.996	.983	.95
8	1.000	1.000	1.000	.999	.996	.98
9	1.000	1.000	1.000	1.000	.999	.99
10	1.000	1.000	1.000	1.000	1.000	.99
11	1.000	1.000	1.000	1.000	1.000	1.00
12	1.000	1.000	1.000	1.000	1.000	1.00
13	1.000	1.000	1.000	1.000	1.000	1.00
14	1.000	1.000	1.000	1.000	1.000	1.00

p								
0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
.000	.000	.000	.00	.000	[illegible]	.000	.000	.000
.005	.000	.000	.00	.000	[illegible]	.000	.000	.000
.027	.004	.000	.00	.000	[illegible]	.000	.000	.000
.091	.018	.002	.00	.000	[illegible]	.000	.000	.000
.217	.059	.009	.00	.000	[illegible]	.000	.000	.000
.403	.151	.034	.00	.001	[illegible]	.000	.000	.000
.610	.304	.095	.01	.004	[illegible]	.000	.000	.000
.787	.500	.213	.05	.017	[illegible]	.000	.000	.000
.905	.696	.390	.13	.057	[illegible]	.000	.000	.000
.966	.849	.597	[illegible]	.148	[illegible]	.002	.000	.000
.991	.941	.783	.48	.314	[illegible]	.013	.001	.000
.998	.982	.909	[illegible]	.539	[illegible]	.056	.005	.000
1.000	.996	.973	.87	.764	[illegible]	.184	.036	.000
1.000	1.000	.995	.96	.920	[illegible]	.451	.171	.010
1.000	1.000	1.000	.99	.987	[illegible]	.794	.537	.140

d. $n = 20$

	0.01	0.05	0.10	0.20	0.25	0.30
0	.818	.358	.122	.012	.003	.001
1	.983	.736	.392	.069	.024	.008
2	.999	.925	.677	.206	.091	.035
3	1.000	.984	.867	.411	.225	.107
4	1.000	.998	.957	.630	.415	.238
5	1.000	1.000	.989	.804	.617	.416
6	1.000	1.000	.998	.913	.786	.608
7	1.000	1.000	1.000	.968	.898	.772
8	1.000	1.000	1.000	.990	.959	.887
9	1.000	1.000	1.000	.997	.986	.952
x 10	1.000	1.000	1.000	.999	.996	.983
11	1.000	1.000	1.000	1.000	.999	.995
12	1.000	1.000	1,000	1.000	1.000	.999
13	1.000	1.000	1.000	1.000	1.000	1.000
14	1.000	1.000	1.000	1.000	1.000	1.000
15	1.000	1.000	1.000	1.000	1.000	1.000
16	1.000	1.000	1.000	1.000	1.000	1.000
17	1.000	1.000	1.000	1.000	1.000	1.000
18	1.000	1.000	1.000	1.000	1.000	1.000
19	1.000	1.000	1.000	1.000	1.000	1.000

p								
0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
.000	.000	.000	.00	.000	[illegible]	.000	.000	.000
.001	.000	.000	.00	.000	[illegible]	.000	.000	.000
.004	.000	.000	.00	.000	[illegible]	.000	.000	.000
.016	.001	.000	.00	.000	[illegible]	.000	.000	.000
.051	.006	.000	.00	.000	[illegible]	.000	.000	.000
.126	.021	.002	.00	.000	[illegible]	.000	.000	.000
.250	.058	.006	.00	.000	[illegible]	.000	.000	.000
.416	.132	.021	.001	.000	[illegible]	.000	.000	.000
.596	.252	.057	.00	.001	[illegible]	.000	.000	.000
.755	.412	.128	.01	.004	[illegible]	.000	.000	.000
.872	.588	.245	.04	.014	[illegible]	.000	.000	.000
.943	.748	.404	.11	.041	[illegible]	.000	.000	.000
.979	.868	.584	.22	.102	[illegible]	.000	.000	.000
.994	.942	.750	.39	.214	[illegible]	.002	.000	.000
.998	.979	.874	.58	.383	[illegible]	.011	.000	.000
1.000	.994	.949	.76	.585	[illegible]	.043	.003	.000
1.000	.999	.984	.89	.775	[illegible]	.133	.016	.000
1.000	1.000	.996	.96	.909	[illegible]	.323	.075	.000
1.000	1.000	.999	.99	.975	[illegible]	.608	.264	.001
1.000	1.000	1.000	.99	.997	[illegible]	.878	.642	[illegible]