STAT 512 Midterm 2 (Total 60+5 Points) – Spring 2014

Name: Solution

Section#3 --- 8:30am; **Section#2 --- 9:30am**

- **Exam time: 8:00-10:00pm.**
- **Must show all work to get credits.**

1. (8 points) Short answer questions.

(1-1) How would you check the constant variance assumption? Give an example of a plot and a formal test.

plot: residual v.s. predictor

test: Levene's test

(1-2) A qualitative variable has 4 possible values, how may indicator variables we need to use to describe it?

3

(1-3) What are the advantages and disadvantages of weighted least squares?

Advantage: ① Improved parameter estimates and CIs

② Valid inference in presence of heteroscedasticity

Disadvantage: No clear interpretation for MSE

MSE will be close to 1 if error variance is modeled well

2. (22 points) A large grocery retailed tracks productivity and costs of its facilities closely. Data were collected from a single distribution center for a certain period. Each data point for each variable represents one week of activity. The variables included are the number of cases shipped (X1), the indirect costs of the total labor hours as percentage (X2), and the total labor hours (Y).

   We fit a model with the main effects of both variables and got the following output:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	[2]	[61423]	[30726.5]	[1.11]	[0.353]
Error	[17]	472033	[27767]		>0.10
Corrected Total	19	533456			

(2-1) Fill in the missing values in the above output.

(2-2) How many observations are used in the analysis?

20

(2-3) Calculate the R-Square and adjusted R-square. Which one is more appropriate to select a model?

$R^2 = \frac{61453}{533456} = 11.52\%$

adjusted $R^2 = 1 - \frac{472033/17}{533456/19} = 1 - \frac{27767}{28077} = 1 - 0.989 = 1.1\%$

adjusted R^2

(2-4) What is the estimate of the variance (σ^2)?

$\hat{\sigma}^2 = 27767$

(2-5) Refer to the following output, using the first observation to calculate and explain the differences between residual, studentized residuals, and studentized deleted residuals.

Model: MODEL1
Dependent Variable: hour

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS		
												Intercept	X1	X2
1	420.0000	425.5993	88.6746	-5.5993	141.1	-0.0397	0.000	-0.0385	0.2832	1.6729	-0.0242	-0.0241	0.0133	0.0188
2	530.0000	367.0058	59.3521	162.9942	155.7	1.047	0.053	1.0500	0.1269	1.1248	0.4002	0.3593	-0.2762	-0.1722
3	550.0000	391.7197	67.2345	158.2803	152.5	1.038	0.070	1.0407	0.1628	1.1771	0.4589	0.4461	-0.2774	-0.2902
4	265.0000	362.6232	52.3784	-97.6232	158.2	-0.617	0.014	-0.6055	0.0988	1.2435	-0.2005	-0.1719	0.1265	0.0751
5	180.0000	317.9857	43.6913	-137.9857	160.8	-0.858	0.018	-0.8511	0.0687	1.1278	-0.2313	-0.0164	0.0423	-0.1081
6	170.0000	380.9618	51.4399	-210.9618	158.5	-1.331	0.062	-1.3643	0.0953	0.9530	-0.4428	-0.4016	0.2028	0.2481
7	130.0000	302.8365	50.2469	-172.8365	158.9	-1.088	0.039	-1.0942	0.0909	1.0626	-0.3460	-0.0099	0.1069	-0.1938
8	304.0000	400.8994	50.0994	-96.8994	158.9	-0.610	0.012	-0.5981	0.0904	1.2340	-0.1885	-0.1345	-0.0118	0.1236
9	380.0000	342.2989	40.3367	37.7011	161.7	0.233	0.001	0.2266	0.0586	1.2619	0.0565	0.0293	-0.0214	0.0013
10	415.0000	387.3510	46.3601	27.6490	160.1	0.173	0.001	0.1677	0.0774	1.2933	0.0486	0.0171	0.0203	-0.0184
11	401.0000	440.7780	90.1635	-39.7780	140.1	-0.284	0.011	-0.2760	0.2928	1.6720	-0.1776	-0.0074	-0.1370	0.0711
12	560.0000	385.7797	112.6709	174.2203	122.8	1.419	0.565	1.4663	0.4572	1.5138	1.3457	-0.6249	1.2690	0.1834
13	190.0000	292.4810	60.7101	-102.4810	155.2	-0.660	0.022	-0.6491	0.1327	1.2793	-0.2539	0.1173	-0.0654	-0.1953
14	270.0000	235.4709	86.6261	34.5291	142.3	0.243	0.007	0.2357	0.2703	1.6267	0.1435	-0.0499	-0.0299	0.1222
15	235.0000	327.9579	46.7243	-92.9579	159.9	-0.581	0.010	-0.5695	0.0786	1.2258	-0.1664	0.0446	-0.0655	-0.0825
16	245.0000	376.5844	40.8323	-131.5844	161.6	-0.814	0.014	-0.8061	0.0600	1.1324	-0.2037	-0.1150	-0.0194	0.0786
17	600.0000	408.8754	54.1417	191.1246	157.6	1.213	0.058	1.2310	0.1056	1.0222	0.4229	0.2212	0.1385	-0.2578
18	320.0000	357.8574	49.5681	-37.8574	159.1	-0.238	0.002	-0.2312	0.0885	1.3028	-0.0720	0.0125	-0.0471	-0.0111
19	710.0000	274.9298	66.2511	435.0702	152.9	2.846	0.507	3.8146	0.1581	0.2046	1.6529	-0.0137	-0.7680	1.0440
20	160.0000	255.0044	75.1144	-95.0044	148.7	-0.639	0.035	-0.6272	0.2032	1.3996	-0.3167	0.1063	0.0503	-0.2636

(1) Residual: $\hat{e}_1 = Y_1 - \hat{Y}_1 = 420 - 425.5993 = -5.5993$

(2) Student Residual: $\hat{e}_i^* = \dfrac{e_i}{\sqrt{MSE \times (1-h_{ii})}} = \dfrac{-5.5993}{\sqrt{27767 \times (1-0.2832)}} = \dfrac{-5.5993}{141.0794} = -0.0397$

(3) Studentized Deleted Residual. $t_i = \dfrac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} = \dfrac{-5.5993}{\sqrt{29500 \times (1-0.2832}} = -0.0385$

145.4133

$(n-p)\,MSE = (n-p-1)\cdot MSE_{(i)} + \dfrac{e_i^2}{1-h_{ii}}$

$MSE_{(i)} = \dfrac{(n-p)\,MSE - e_i^2/(1-h_{ii})}{n-p-1} = \dfrac{17 \times 27767 - \frac{5.5993^2}{1-0.2832}}{20-3-1} = \dfrac{472039 - 43.7391}{16} = \dfrac{472000}{16}$

$= 29500$

(2-6) Is there any outlier based on a formal test? Give the test statistic, the critical value, and your conclusion.

$\alpha = 0.05$ $\quad n = 20$

$t_{n-p-1}\left(1-\frac{\alpha}{2n}\right) = t_{16}\left(1-\frac{0.05}{2\times20}\right) = t_{16}(1-0.0013) \approx t_{16}(0.999) = 3.686$

Yes. Studentized deleted residual $\Rightarrow$ #19 is an outlier.

(2-7) Is there any observation having substantial influence on the regression coefficients?

$F_{3,17}(0.5)$

h_{ii}: $\frac{2p}{n} = \frac{2\times3}{20} = 0.3$ $\quad$ #12

DEFITS: >1 (small, medium data) $\quad$ #12, #19

DFBETAS: >1 (small-medium datasets): #12 $\rightarrow$ affect X_1

#19 $\rightarrow$ affect X_2

(2-8) Based on the results of (2-6) and (2-7), discuss the next step analysis.

3. (30+5 points) An experiment was conducted to test the effect of fertilizer on lettuce production. Five different types of fertilizer were applied to four replicated plots in the experiment. The data are the number of heads of lettuce harvested from the plot.

(3-1) Write down the cell means model and corresponding assumptions.

$Y_{ij} = \mu_i + \varepsilon_{ij}$ $\quad i = 1, 2, 3, 4, 5$

$\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

(3-2) What are the parameters in the cell means model in (3-1)?

$\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \quad \sigma^2$

(3-3) Write down the factor effects model and assumptions.

$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ $\quad \varepsilon_{ij} \sim N(0, \sigma^2)$

$\Sigma \tau_i = 0$

(3-4) What are the parameters in the factor effects model?

$\mu, \quad \tau_1, \tau_2, \tau_3, \tau_4, \tau_5, \quad \sigma^2$

(3-5) Fill in the 3 missing values in the brackets of the following SAS output.

SAS code:

```
PROC MEANS data=exam2Q3 mean std stderr clm;
      CLASS type;
      VAR production;
RUN;
```

SAS output:

The MEANS Procedure

Analysis Variable : production

type	N Obs	Mean	Std Dev	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean
1	4	11.0000000	2.1602469	[1.0801]	[ ]	[ ]

SE: $\frac{2.1602}{\sqrt{4}} = 1.0801$ $\quad t_{3,0.05} = 3.182$

Lower: $11 - 3.182 \times 1.0801 = 11 - 3.4369 = 7.5631$

Upper: $11 + 3.182 \times 1.0801 = 11 + 3.4369 = 14.4369$

(3-6) Using GLM procedure, we have the following output. Fill in the 2 missing values in the brackets of the following output.

2-sided area

$t_{15,0.05} = 2.131$

Lower: $11 - 2.131 \times \sqrt{2.15/4} = 11 - 1.5623 = 9.4377$

Upper: $11 + 1.5623 = 12.5623$

The GLM Procedure
t Confidence Intervals for production

Alpha	0.05
Error Degrees of Freedom	15
Error Mean Square	2.15

type	N	Mean	95% Confidence Limits	
1	4	11.0000	[9.4377]	[12.5623]
2	4	14.2500		
3	4	15.0000		
4	4	15.7500		
5	4	14.7500		

(3-7) Compare the CIs in (3-5) and (3-6) and interpret it.

(3-6) is narrower b/c using more df to estimate variance.

(3-8) Construct simultaneously CIs for all 5 types using Bonferroni method based on the output from PROC GLM in (3-6).

$t_{15,\ 0.05/5} = t_{15,\ 0.01} = 2.947$

half width of CI $= 2.947 \times \sqrt{\frac{2.15}{4}} = 2.1606$

type 1: $11 - 2.1606,\quad 11 + 2.1606 = (\quad)$

2: $14.25 - 2.1606 \quad 14.25 + 2.1606$

3: $15 - 2.1606 \quad 15 + 2.1606$

4: $15.75 - 2.1606 \quad 15.75 + 2.1606$

5: $14.75 - 2.1606 \quad 14.75 + 2.1606$

(3-9) If the type 1 fertilizer is the current standard and type 2-5 are the modified fertilizers, we would like to compare the average of the modified ones with the current standard fertilizer. How would you set up the contrast statement in SAS?

Contrast '1 v.s 2.3.45" type 1 $-\frac{1}{4}$ $-\frac{1}{4}$ $-\frac{1}{4}$ $-\frac{1}{4}$;

(3-10) What is the point estimate of the contrast in (3-9)?

$$\hat{C}_1 = \bar{Y}_1 - \frac{\bar{Y}_2+\bar{Y}_3+\bar{Y}_4+\bar{Y}_5}{4}$$

$$= 11 - \frac{14.25+15+15.75+14.75}{4}$$

$$= 11 - 14.9375$$

$$= -3.9375$$

(3-11) Bonus question. (+5). Perform the hypothesis test that the contrast in (3-9) equals to -2?

$$H_0: C_1 = \mu_1 - \frac{\mu_2+\mu_3+\mu_4+\mu_5}{4} = -2$$

$$H_a: C_1 = \mu_1 - \frac{\mu_2+\mu_3+\mu_4+\mu_5}{4} \neq -2$$

TS: $\hat{C}_1 = -3.9375$

$$Var(\hat{C}_1) = Var\left(\bar{Y}_1 - \frac{\bar{Y}_2+\bar{Y}_3+\bar{Y}_4+\bar{Y}_5}{4}\right)$$

$$= Var(\bar{Y}_1) + \frac{1}{16} Var\left(\sum_{i=2}^{5} \bar{Y}_i\right) \qquad Var(\bar{Y}_i) = \frac{\sigma^2}{n} = \frac{\sigma^2}{4}$$

$$= \frac{\sigma^2}{4} + \frac{1}{16} \times 4 \times \frac{\sigma^2}{4}$$

$$= \frac{5}{16}\sigma^2$$

$$SE(\hat{C}_1) = \sqrt{\frac{5}{16}\hat{\sigma}^2} = \sqrt{\frac{5}{16} \times 2.15} = 0.8197$$

$$TS = \frac{-3.9375-(-2)}{0.8197} = \frac{-1.9375}{0.8197} = -2.3637$$

$$t_{15,\,0.05} = -2.1314$$

Reject H_0 and conclude that the contrast does not equal to -2.