STOR 455 Final **December 16, 2010**

Name: ________________ **PID:______________**

INSTRUCTIONS:
BOTH THE EXAM AND THE BUBBLE SHEET WILL BE COLLECTED. YOU MUST PRINT YOUR NAME AND SIGN THE HONOR PLEDGE ON THE BUBBLE SHEET. YOU MUST BUBBLE-IN YOUR NAME & YOUR STUDENT IDENTIFICATION NUMBER.

EACH QUESTION HAS ONLY ONE CORRECT CHOICE (decimals may need rounding).

USE "NUMBER 2" PENCIL ONLY - DO NOT USE INK - FILL BUBBLE COMPLETELY.

NO NOTES OR REMARKS ARE ACCEPTED - DO NOT TEAR OR FOLD THE BUBBLE SHEET.

A GRADE OF ZERO WILL BE ASSIGNED FOR THE ENTIRE EXAM IF THE BUBBLE SHEET IS NOT FILLED OUT ACCORDING TO THE ABOVE INSTRUCTIONS.

QUESTIONS are worth **1 point** each.

Questions 1-2 are based on the following
A data set "blah" contains response variable Y and predictor variables X_1 and X_2. It is known from the scientific background of the experiment that the appropriate model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \sqrt{X_1^2 + X_2^2} \cdot \xi$ where ξ are independent $N(0,\sigma^2)$.

1. The appropriate method for fitting this model is
A) Box-Cox transformation
B) This model cannot be fitted
C) Ordinary Least Squares
D) Weighted Least Squares
E) None of the above

2. The appropriate SAS code is

A)
```
proc reg data=blah;
model Y=X1 X2;
run;
```

B)
```
proc transreg data=blah; W;
model boxcox(Y)=identity(X1 X2);
run;
```

C)
```
data blah; set blah;
W=X1*X1+X2*X2;
run;
proc reg data=blah;
weight W;
model Y=X1 X2;
run;
```

D)
```
data blah; set blah;
W=1/(X1*X1+X2*X2);
run;
proc reg data=blah;
weight W;
model Y=X1 X2;
run;
```

E) None of the above

3. Which of the following functions is linear in unknown parameters (symbols β)?

$A)(\beta_1 x_1 + \beta_2 x_2)^2 \quad B)\beta_0 + \sin(\beta_1 x) \quad C)e^{\beta_0+\beta_1 x_1} \quad D)\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

E) None of the above

4. Which of the following functions **cannot** be made into a function linear in unknown parameters (symbols β) using a Box-Cox Transformation?

$A)(\beta_1 x_1 + \beta_2 x_2)^2 \quad B)\beta_0 + \sin(\beta_1 x) \quad C)e^{\beta_0+\beta_1 x_1} \quad D)\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

E) None of the above

Use the following to answer questions 5 - 9:
A researcher wants to evaluate a new methodology in determining a chemical concentration of a particular heavy metal in soil. To investigate the relationship between the true and measured concentration, the researcher makes 32 samples containing a known (preselected) amount of the heavy metal. These samples are then analyzed using a technician who is unaware of the true concentration of the heavy metal.

5. The target population of items in this study is
A) measured concentration B) all soil samples C) both B, D
D) soil samples prepared in the lab E) none of the above

6. The study population of items in this study is
A) measured concentration B) all soil samples C) both B, D
D) soil samples prepared in the lab E) none of the above

7. The response variable in this study is
A) true concentration B) measured concentration C) not enough info
D) soil sample E) None of the above

8. The predictor variable in this study is
A) true concentration B) measured concentration C) not enough info
D) soil sample E) None of the above

9. The equation of the least-squares regression line is

$$\hat{y} = -0.1046 + 0.9877 \cdot x$$

Which of the following descriptions of the value of the slope is the correct description?
A) The measured concentration is expected to decrease by 0.1046 when the true concentration increases by 1.
B) The measured concentration is expected to increase by 0.9877 when the true concentration increases by 1.
C) We cannot interpret the slope because we cannot have a negative concentration.
D) None of the above

Use the following to answer questions 10 - 18:
Below is a small data set, and we want to fit the linear regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \xi$. In matrix notation this model can be written as $Y = \mathbf{X}\beta + \xi$.

Y	19	13	10	3	6
X1	1	-1	0	-1	1
X2	-2	-1	0	1	2

10. The matrix **X** is

$$A)\ \mathbf{X} = \begin{pmatrix} 1 & -2 \\ -1 & -1 \\ 0 & 0 \\ -1 & 1 \\ 1 & 2 \end{pmatrix} \quad B)\ \mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 0 & 1 & -1 \\ -2 & -1 & 0 & 1 & 2 \end{pmatrix} \quad C)\ \mathbf{X} = \begin{pmatrix} 1 & 1 & -2 \\ 1 & -1 & -1 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \\ 1 & 1 & 2 \end{pmatrix} \quad D)\ \mathbf{X} = \begin{pmatrix} 19 & 1 & -2 \\ 13 & -1 & -1 \\ 10 & 0 & 0 \\ 3 & -1 & 1 \\ 6 & 1 & 2 \end{pmatrix}$$

E) None of the above

11. Find the vector b, the LS estimator of β.

$$A)\ b = \begin{pmatrix} 2.25 \\ -3.6 \end{pmatrix} \quad B)\ b = \begin{pmatrix} 255 \\ 36 \\ -360 \end{pmatrix} \quad C)\ b = \begin{pmatrix} 10.2 \\ 2.25 \\ -3.6 \end{pmatrix} \quad D)\ b = \begin{pmatrix} 51 \\ 9 \\ -36 \end{pmatrix}$$

E) None of the above

12. Compute the SSE
A) 50 B) 150 C) 5 D) 0 E) None of the above is within ±1

13. The number of degrees of freedom in MSE is
A) 2 B) 0 C) 4 D) 3 E) None of the above

14. Compute the SSR (sometimes also called SSM)
A) 50 B) 150 C) 5 D) 0 E) None of the above is within ±1

15. What is the number of degrees of freedom in MSR (sometimes also called MSM)
A) 2 B) 0 C) 4 D) 3 E) None of the above

16. Which of the following can be used for testing H_0: $\beta_1=0$, $\beta_2=0$?
A) MSR/MSE B) MSTO/MSE C) SSM/SSTO D) SSR/SSE
E) None of the above

17. Which of the following defines R^2?
A) MSR/MSE B) MSTO/MSE C) SSR/SSTO D) SSR/SSE
E) None of the above

18. Which of the following defines adjusted R^2?
A) MSR/MSE B) MSTO/MSE C) SSR/SSTO D) SSR/SSE
E) None of the above

Use the following to answer questions 19 – 26:
Crime-related and demographic statistics for 47 US states in 1960 were collected from the FBI's Uniform Crime Report and other government agencies to determine how the variable crime rate depends on the other variables measured in the study. Following is a description of the variables:

```
R: Crime rate: # of offenses reported to police per million population
Age: The number of males of age 14-24 per 1000 population
S: Indicator variable for Southern states (0 = No, 1 = Yes)
Ed: Mean # of years of schooling x 10 for persons of age 25 or older
Ex0: 1960 per capita expenditure on police by state and local government
Ex1: 1959 per capita expenditure on police by state and local government
LF: Labor force participation rate per 1000 civilian urban males age 14-24
M: The number of males per 1000 females
N: State population size in hundred thousands
NW: The number of non-whites per 1000 population
U1: Unemployment rate of urban males per 1000 of age 14-24
U2: Unemployment rate of urban males per 1000 of age 35-39
W: Median value of transferable goods and assets or family income
X: The number of families per 1000 earning below 1/2 the median income
```

Here is the SAS output obtained by fitting a linear regression model with R as response variable and all the other variables as explanatory variables.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-691.83759	155.88792	-4.44	<.0001	0
Age	1	1.03981	0.42271	2.46	0.0193	2.69802
S	1	-8.30831	14.91159	-0.56	0.5812	4.87675
Ed	1	1.80160	0.64965	2.77	0.0091	5.04944
Ex0	1	1.60782	1.05867	1.52	0.1384	94.63312
Ex1	1	-0.66726	1.14877	-0.58	0.5653	98.63723
LF	1	-0.04103	0.15348	-0.27	0.7909	3.67756
M	1	0.16479	0.20993	0.78	0.4381	3.65844
N	1	-0.04128	0.12952	-0.32	0.7520	2.32433
NW	1	0.00717	0.06387	0.11	0.9112	4.12327
U1	1	-0.60168	0.43715	-1.38	0.1780	5.93826
U2	1	1.79226	0.85611	2.09	0.0441	4.99762
W	1	0.13736	0.10583	1.30	0.2033	9.96896
X	1	0.79293	0.23509	3.37	0.0019	8.40945

19. Which variables have a multicollinearity problem?
A) Intercept B) Ex0 C) Ex1 D) Both B and C
E) None of the above

20. When running a backward selection, which of the variables would get dropped from the model first?
A) X B) Ex1 C) NW D) Not enough info E) None of the above

Use these additional SAS outputs to answer questions 21 – 26:
After model selection we include only five variables in our model: Age, Ed, Ex0, U2 and X. Below is a SAS output from fitting this model.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	?	?????	10041	22.13	<.0001
Error	??	?????	????		
Corrected Total	46	68809			

Parameter Estimates

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	-524.37433	95.11557	-5.51	<.0001	385014	13791
Age	1.01982	0.35320	2.89	0.0062	550.83962	3782.81167
Ed	2.03077	0.47419	4.28	0.0001	7259.67450	8322.11780
Ex0	1.23312	0.14163	8.71	<.0001	31739	34394
U2	0.91361	0.43409	2.10	0.0415	2173.86286	2009.88258
X	0.63493	0.14685	4.32	<.0001	8482.73176	8482.73176

Obs	RStudent	Hat Diag H	Cov Ratio	DFFITS
1	-0.4906	0.0990	1.2416	-0.1626
2	1.4616	0.0782	0.9207	0.4257
3	0.6046	0.1093	1.2328	0.2118
4	0.5994	0.2075	1.3869	0.3068
5	-0.2449	0.1367	1.3313	-0.0974
6	-0.3181	0.1777	1.3891	-0.1479
7	1.1278	0.0579	1.0202	0.2797
8	1.0086	0.1031	1.1121	0.3419
9	0.4704	0.0974	1.2430	0.1545
10	-0.3639	0.0865	1.2446	-0.1120
11	3.5227	0.1000	0.2547	1.1741
12	1.1253	0.0506	1.0131	0.2598
13	-1.0518	0.1363	1.1400	-0.4178
14	-0.3058	0.0974	1.2670	-0.1005
15	-0.2321	0.1139	1.2982	-0.0832
16	-0.2079	0.1228	1.3135	-0.0778
17	-0.4763	0.1022	1.2486	-0.1607
18	-0.9402	0.0929	1.1214	-0.3010
19	-2.2305	0.0852	0.6274	-0.6807
20	0.1226	0.2143	1.4727	0.0640
21	0.2674	0.0744	1.2395	0.0758
22	-1.8952	0.1148	0.7821	-0.6825
23	1.8350	0.1147	0.8066	0.6606
24	0.2800	0.1052	1.2809	0.0960
25	-0.6461	0.1485	1.2798	-0.2698
26	0.6868	0.2100	1.3683	0.3541
27	0.3190	0.2107	1.4471	0.1649
28	0.0152	0.1082	1.3004	0.0053
29	-2.6667	0.2678	0.5934	-1.6129
30	-0.0913	0.1569	1.3739	-0.0394
31	-0.0647	0.1138	1.3078	-0.0232
32	-0.1360	0.0767	1.2525	-0.0392
33	1.1263	0.0823	1.0478	0.3373
34	-0.4045	0.0601	1.2039	-0.1023
35	-0.4308	0.1564	1.3370	-0.1855
36	1.5850	0.2216	1.0338	0.8456
37	-0.4326	0.2457	1.4950	-0.2469
38	0.1841	0.1198	1.3108	0.0679
39	0.4727	0.0800	1.2191	0.1394
40	0.2967	0.0594	1.2168	0.0746
41	0.7949	0.2013	1.3217	0.3991
42	0.1302	0.1231	1.3191	0.0488
43	-0.9433	0.1122	1.1447	-0.3353
44	-0.5318	0.0981	1.2326	-0.1753
45	-0.7245	0.2062	1.3510	-0.3692
46	-1.3750	0.1543	1.0394	-0.5873
47	-0.9000	0.1089	1.1539	-0.3146

21. What is the estimate of σ?
A) 22.13 B) 100.2 C) 453.75 D) 21.3 E) None of the above

22. Find the extra sum of squares SSR(U2|Age, Ed, Ex0)
A) 2009.9 B) 2173.9 C) 2.1 D) 68809 E) None of the above

23. Find the extra sum of squares SSR(U2|Age, Ed, Ex0, X)
A) 2009.9 B) 2173.9 C) 2.1 D) 68809 E) None of the above

24. Which statistic in the SAS output above can be used to detect outliers in the response variable?
A) RStudent B) Hat Diag C) Cov Ratio D) DFFITS E) None of the above

25. Which statistic in the SAS output above can be used to detect outliers in the explanatory variable?
A) RStudent B) Hat Diag C) Cov Ratio D) DFFITS E) None of the above

26. Are there any influential observations? Why?
A) Yes, because observation 11 has |RStudent|>3.
B) Yes, because observations 11 and 29 have |DFFITS|>1
C) No, there is no observation with Hat Diag>0.5
D) No because there are no outliers in the data set.
E) None of the above

Use the following to answer questions 27 – 29:
27. Assume that we are fitting a simple linear regression with non-negative response variable Y and predictor X in a data set dog. Residual analysis leads you to believe that a transformation of Y might be needed. What SAS code gives you an estimate of λ in the Box-Cox transformation?

A)
```
proc transreg data=dog;
boxcox(X)=identity(Y);
run;
```

B)
```
proc reg data=dog;
Y=X/selection=boxcox;
run;
```

C)
```
proc transreg data=dog;
boxcox(X)=Y;
run;
```

D)
```
proc transreg data=dog;
boxcox(Y)=identity(X);
run;
```

E) None of the above

28. If the Box-Cox procedure selected λ=0, what model is it suggesting?
A) $Y= (\beta_0+\beta_1 X_1)^2+\xi$ B) $\sqrt{Y}= \beta_0+\beta_1 X_1+\xi$ C) $Y^0= \beta_0+\beta_1 X_1+\xi$
D) $\log Y= \beta_0+\beta_1 X_1+\xi$ E) None of the above

29. If the Box-Cox procedure selected λ=1/2, what model is it suggesting?
A) $Y= (\beta_0+\beta_1 X_1)^2+\xi$ B) $\sqrt{Y}= \beta_0+\beta_1 X_1+\xi$ C) $Y^0= \beta_0+\beta_1 X_1+\xi$
D) $\log Y= \beta_0+\beta_1 X_1+\xi$ E) None of the above

Use the following to answer questions 30 – 33:
A researcher studies a relationship between latitude/longitude of a city and the January minimum temperature. The data set contains the normal average January minimum temperature (in Fahrenheit) and longitude and latitude of 56 cities located in the continental USA (lower 48 states). Following is the description of the variables:

```
Temp: Average January minimum temperature in degrees F. from 1931-1960
Lat: Latitude in degrees north of the equator
Long: Longitude in degrees west of the prime meridian
```

We first fit a linear regression model without any interactions. A SAS output follows

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7297.33488	3648.66744	75.88	<.0001
Error	53	2548.64727	48.08768		
Corrected Total	55	9845.98214			

Root MSE	6.93453	R-Square	0.7411
Dependent Mean	26.51786	Adj R-Sq	0.7314
Coeff Var	26.15041		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	110.83077	6.90937	16.04	<.0001
lat	1	-2.16355	0.17570	-12.31	<.0001
long	1	0.13396	0.06314	2.12	0.0386

30. Raleigh has (Lat, Long)=(35.8,78.6) and Los Angeles has (Lat, Long) = (34.1,118.3). Predict the average difference (Raleigh-LA) in Temp between these two cities.
A) 86.1 B) -9.0 C) 43.9 D) -1.6 E) None of the above

31. Honolulu has (Lat, Long)=(21.3,157.8). Is it appropriate to use the above model to predict Honolulu's average January minimum temperature? Why?
A) Yes, the prediction is 85.9
B) No. The longitude is only borderline significant.
C) No. We should not extrapolate this far outside of the predictor values in the data set.
D) Not enough info.
E) None of the above.

We suspect that the relationship between Temp and Long is nonlinear and polynomial regression model might be needed. To determine the degree of the polynomial a model selection procedure is used and the following is the output. (l2 =long*long, l3=long*long*long, etc.)

	R-Square	C(p)	MSE	Variables in Model
1	0.7192	213.4412	51.20572	lat
1	0.0375	857.7697	175.50181	l5
1	0.0304	864.4348	176.78757	l3
1	0.0155	878.5639	179.51317	l4
1	0.0010	892.2157	182.14671	l2
2	0.8920	52.0766	20.06292	lat l4
2	0.8809	62.5808	22.12751	lat l5
2	0.8654	77.2083	25.00250	lat l2
2	0.8415	99.8352	29.44977	lat l3
2	0.7411	194.6616	48.08768	lat long
3	0.9457	3.2876	10.27432	lat long l3
3	0.9289	19.1562	13.45323	lat long l5
3	0.9193	28.2423	15.27343	lat long l4
3	0.9016	45.0485	18.64018	lat l3 l5
3	0.8977	48.7190	19.37548	lat l3 l4
4	0.9467	4.3319	10.28057	lat long l3 l5
4	0.9461	4.9278	10.40229	lat long l2 l3
4	0.9458	5.2566	10.46944	lat long l3 l4
4	0.9310	19.1810	13.31358	lat long l2 l5
4	0.9297	20.4269	13.56806	lat long l4 l5
5	0.9477	5.4129	10.29472	lat long l3 l4 l5
5	0.9473	5.8336	10.38236	lat long l2 l3 l5
5	0.9464	6.6341	10.54915	lat long l2 l3 l4
5	0.9326	19.6755	13.26619	lat long l2 l4 l5
5	0.9066	44.2595	18.38804	lat l2 l3 l4 l5
6	0.9482	7.0000	10.41703	lat long l2 l3 l4 l5

32. Based strictly on Mallow's C(p), which model would you recommend using?
A) temp=lat long l3 B) temp=lat C) temp=lat long l2 l3 l4 l5
D) temp=lat long E) None of the above

33. Is the all-subsets procedure better than the step-wise procedure for this problem?
A) No. Stepwise procedure is always better.
B) Yes. If feasible, the all-subset procedure is always better.
C) It does not matter which model selection procedure we use.

34. A researcher is studying effects of predictors U and V on a response variable Z. It is expected that large values of U will reinforce the effect of V on the response variable Z. What model statement is appropriate for this situation? (V2=V*V, UV=U*V, UZ=U*Z, etc.)

A) model V=U Z UZ; B) model Z=U V; C) model U=V V2;
D) model Z=U V UV; E) linear regression is not able to handle this situation

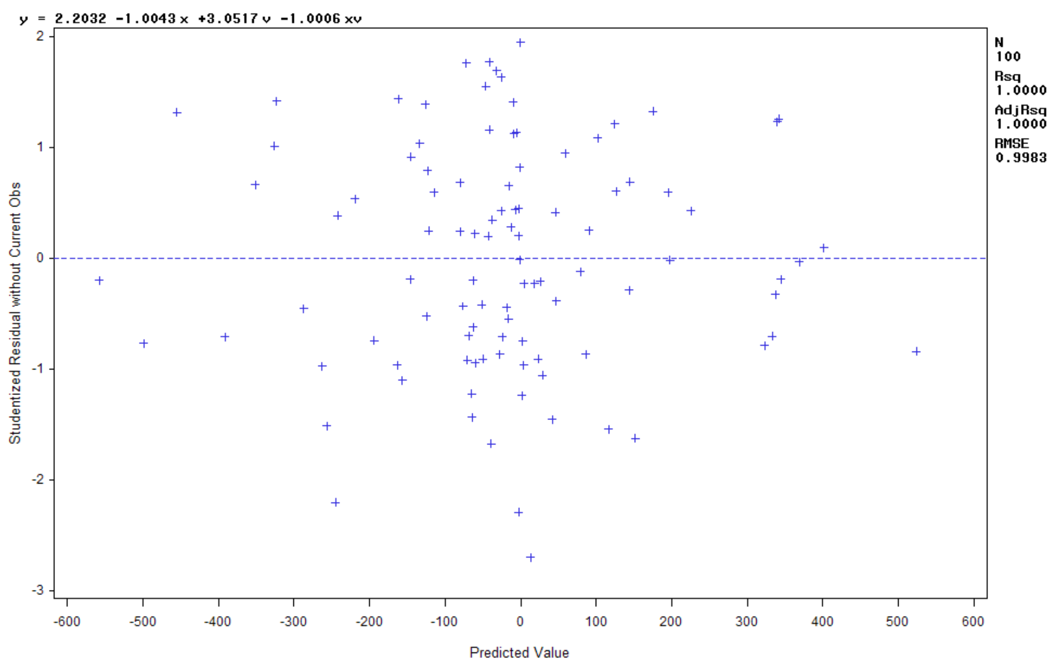
35. What should you do first when you receive a new data set for analysis?

A) Investigate residuals for outliers.
B) Plot the data and investigate all the variables.
C) Fit the largest possible model available
D) Run a stepwise selection procedure.
E) Investigate DFFITS for influential observations.

36. What do we hope to capture within a confidence interval?

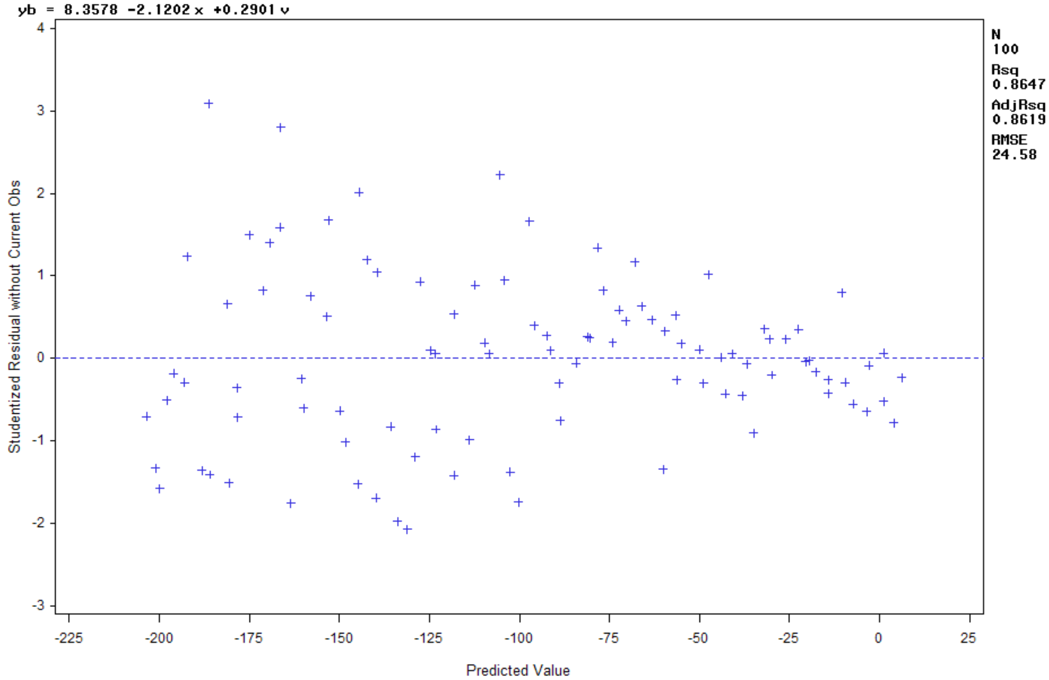
A) The unknown parameter value. B) The sample size.
C) The unknown confidence level. D) The parameter estimate.
E) None of the above.

37. What can you conclude from the following deleted residual plot?



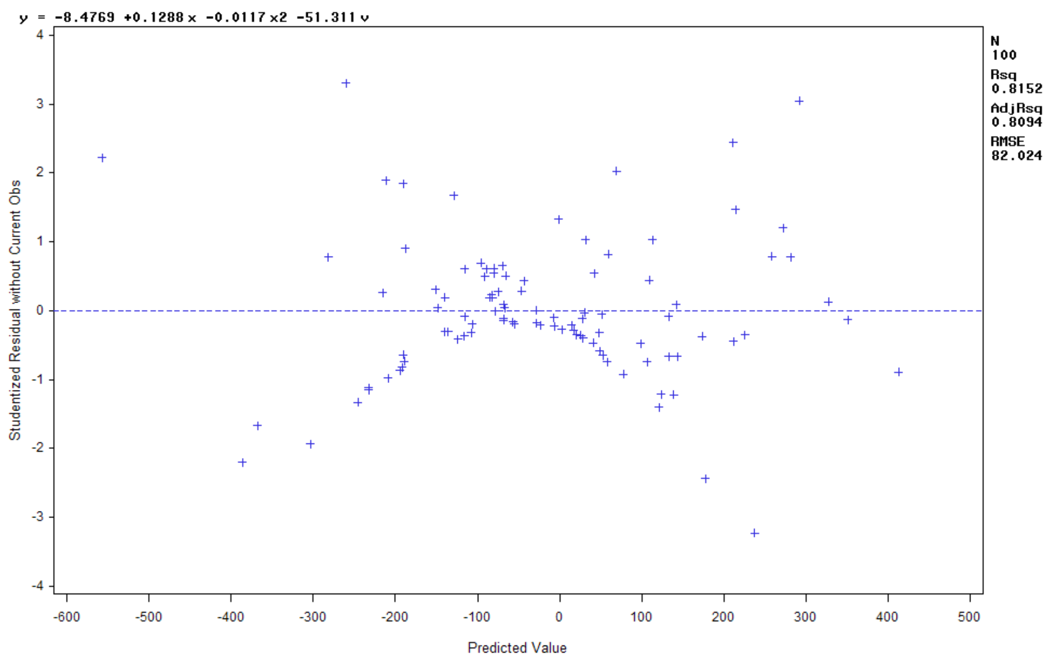
A) The model does not fit B) There are influential observations
C) The model seems to fit well D) The variance of the residuals is decreasing
E) None of the above

38. What can you conclude from the following deleted residual plot?



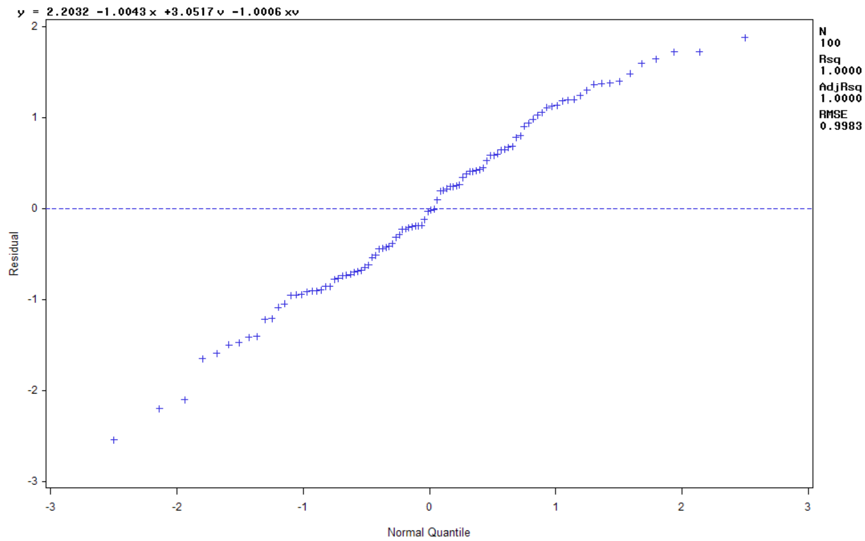
A) The model does not fit
B) There are influential observations
C) The model seems to fit well
D) The variance of the residuals is decreasing
E) None of the above

39. What can you conclude from the following deleted residual plot?



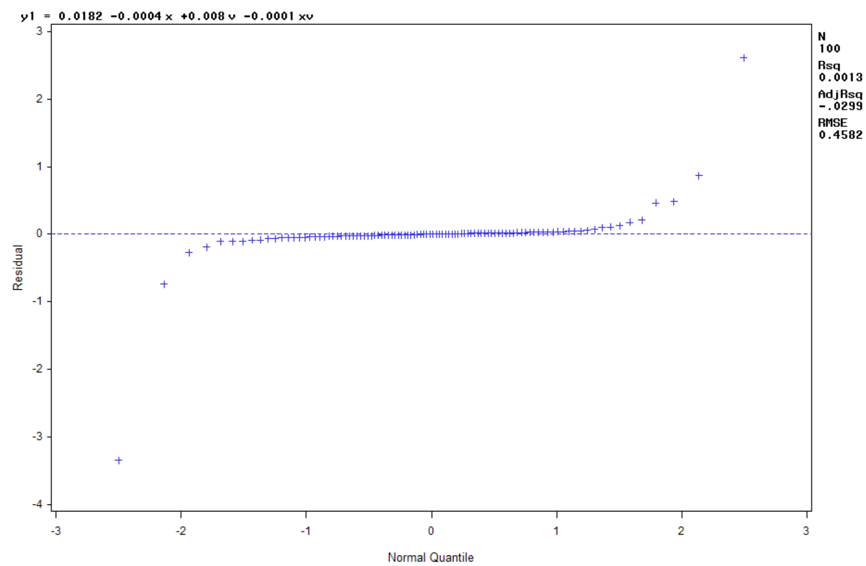
A) The model does not fit
B) There are influential observations
C) The model seems to fit well
D) The variance of the residuals is decreasing
E) None of the above

40. What can you conclude from the following QQ plot?



A) There is nothing we can learn here
B) QQ plot should never be examined, instead examine the residual plot
C) The assumption that residuals are normal appears to be violated
D) The assumption that residuals are normal appears to be valid
E) None of the above

41. What can you conclude from the following deleted QQ plot?



A) There is nothing we can learn here
B) QQ plot should never be examined, instead examine the residual plot
C) The assumption that residuals are normal appears to be violated
D) The assumption that residuals are normal appears to be valid
E) None of the above