Statistics 621 Waiver Exam

August 25, 2002

This is an **open-book, open-notes** exam. You are limited to TWO textbooks, please. No laptops or other computers are allowed.

You have **two hours** for the exam.

The **computer output** associated with one or more items should be considered an essential part of the questions.

The multiple-choice questions are equally weighted. Your score is the total number of correct answers given in questions that are scored. Some questions may be dropped and not counted as part of the overall score.

Please also note the following when filling in the answer sheet.

– Fill in your name and student id number on the answer form.

– Mark the "bubbles" under your name and student id number on the form.

– Choose **one *best* answer** by marking the item on the answer form.

– Mark the answer form using only a **#2 pencil**. Erase all changes completely.

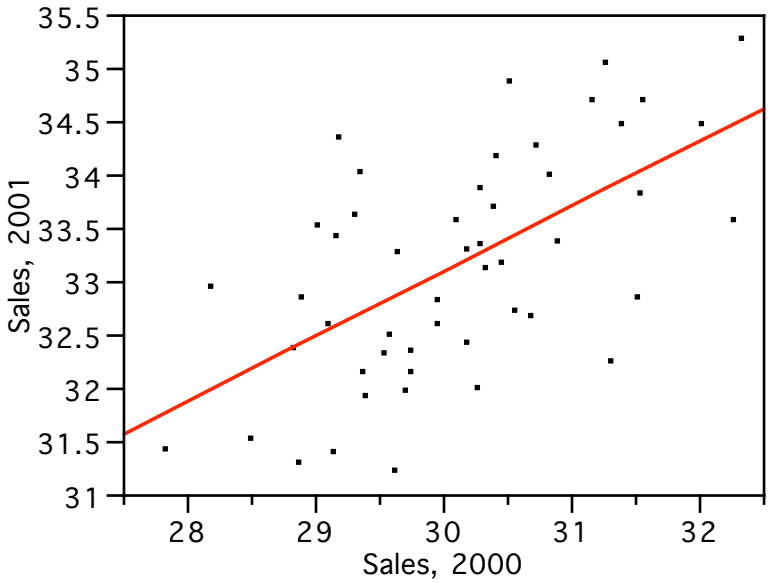
Turn in the solution page only; keep the test. Mark your copy of the exam in order to remember your choices so that you can check how well you did.

Solutions will be posted in WebCafe. A list of those passing the exam will be available from the MBA program office on the third floor of Huntsman Hall.

STOP

Do not turn the page until you are instructed.

(Questions 1– 11) A retail chain has compared the sales at a sample of 50 of its retail stores in 2001 to sales in the same stores during the prior year. Each observation is the typical daily sales at a retail store belonging to this chain in both years. The values are in thousands of US dollars. A least squares regression analysis of the sales in 2001 on the sales in 2000 gives the following results.



Sales, 2001 = 14.8 + 0.61 Sales, 2000

Summary of Fit

RSquare	0.364
Root Mean Square Error	0.844
Mean of Response	33.154
Observations	50.000

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	14.83	3.50	4.23	0.0001
Sales, 2000	0.61	0.12	5.24	<.0001

(1) The shown results imply that the average levels of sales in stores are
- a) About the same in both years.
- a) Significantly different on average, with higher sales in 2000.
- b) Significantly different on average, with higher sales in 2001.
- c) Moving significantly closer together.
- d) Moving significantly farther apart.

(2) The confidence interval for the slope of the model implies that (assuming that the usual assumptions are valid for this model)

a) The population slope is zero.
b) The population slope is not zero.
c) There is a 95% chance that the true slope is between 0.37 and 0.85.
d) There is a 95% chance that the true slope is between 0.49 and 0.73.
e) There is a 12% chance that the slope in the population is larger than 0.61.

(3) According to the fitted model, the 2001 sales in a store that sold $30,000 daily in 2000 would be

a) $14,830
b) $18,300
c) $30,000
d) $33,100
e) $33,500

(4) Managers at two of the stores, call them A and B, are very competitive. In 2000, sales were $32,000 daily in Store A and $30,000 in Store B. According to the fitted model, the estimated difference in sales of these two stores in 2001 is about

a) $0
b) $600
c) $1,200
d) $1,480
e) $2,000

(5) Another store (not included in the analysis) had the following pattern of sales: $29,000 in 2000 rising to $33,000 in 2001. This model suggests that this change is

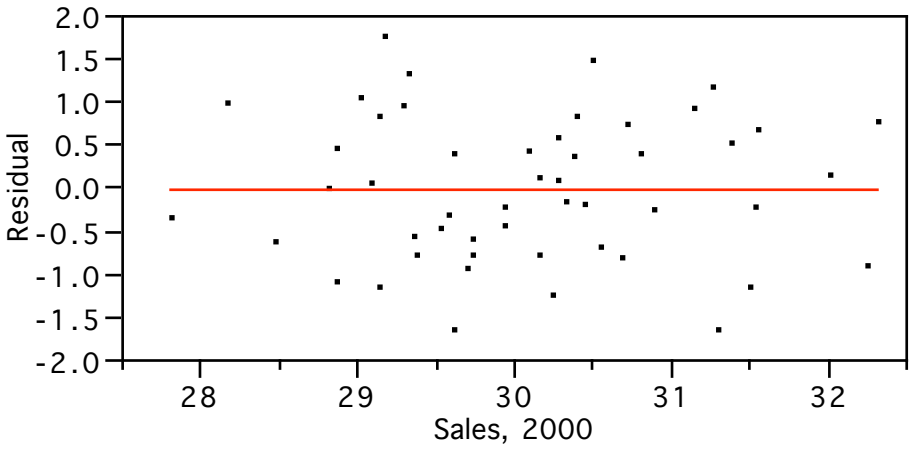
a) Lower than the sales gain that would be predicted.
b) About what would be predicted for the typical outlet.
c) Above what would be predicted, but not very much.
d) So far above the prediction that the manager deserves a bonus.
e) So far below the prediction that the manager should be retrained.

(6) If two stores have the same sales in 2000, what is the approximate probability that they will differ by at least $1,200 in 2001? (i.e., that the absolute size of the difference in sales in 2001 will be larger than $1200) You may assume the standard assumptions for this calculation.

a) About 2.5%
b) About 5%
c) About 1/6
d) About 1/3
e) About 1/2

(7) Data for which of the following types of stores would lead to the most improvement in the length of the confidence interval (i.e. shorten) for the slope in this problem?

a) A store with daily sales of $27,000 in 2000.
b) A store with daily sales of $30,000 in 2000.
c) A store with daily sales of $30,000 in 2001.
d) A store with daily sales of $33,000 in 2001.
e) Any of the above would offer the same improvement.



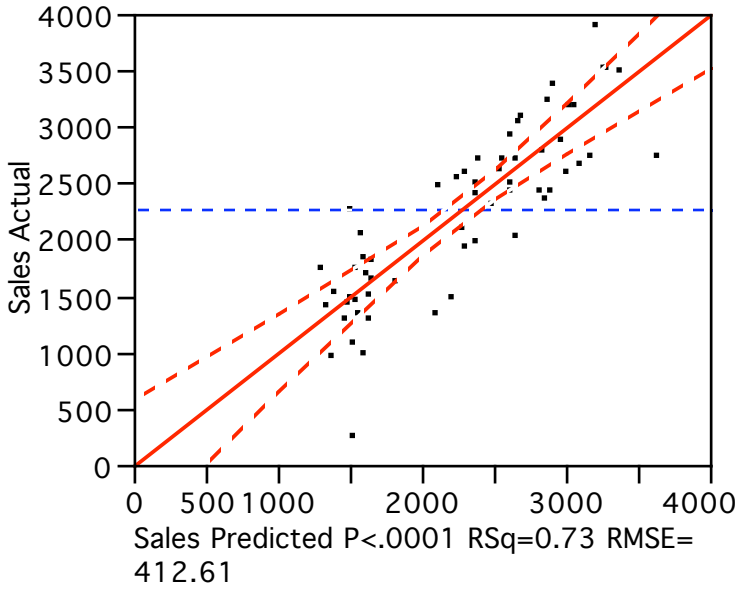
(8) The residual plot shown above for the fitted model indicates that
- a) The model has omitted an important predictor of sales.
- b) The data are dependent.
- c) The variance of the underlying error terms is not constant.
- d) The distribution of the error terms is not normal.
- e) The data are consistent with the usual assumptions of regression.

(9) It was later learned that 25 of the stores in this analysis were on the west coast of the US, and 25 of the stores were located on the east coast of the US. To handle this information, one could
- a) Divide the data into two groups to be analyzed separately.
- b) Add a categorical predictor indicating location to the regression.
- c) Color-code the data to identify unusual patterns associated with one group.
- d) Any of the above would be useful.
- e) None of the above; leave the model alone.

(10) An analyst reported that the Durbin-Watson statistic for this analysis was 0.75. This result implies that
- a) The model is not valid; the data are not normally distributed.
- b) The model is not valid; the data lack constant variance.
- c) The model is not valid; the data are dependent.
- d) The model is not valid; an important time series factor has been omitted.
- e) The data have a curious pattern in the errors that deserves further study.

(11) A consultant advised management that, rather than use regression, it should instead use a paired t-test to assess the overall change in store sales. This advice
- a) Is not useful; a t-test is a special case of regression.
- b) Is not useful; the intercept of the model captures the average difference.
- c) Is not useful; such a method would not find a significant effect.
- d) Is useful and gives an important alternative analysis.
- e) Is useful but will be more difficult to interpret without the regression slope.

(Questions 12-21) A vendor operates a collection of lunch carts in Philadelphia. The vendor would like to understand how the number of beverages sold (soda, iced tea, etc) depends on the weather. Over the last 60 weekdays, the vendor has collected data on the number of carts operating that day as well as the high temperature and whether it rained or not. The response (Sales) is the number of beverage "servings" sold that day (for example, 1 bottle of Coke = 1 beverage sale). The following regression model summarizes one analysis of this data.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	25225403	6306351	37.0421
Error	55	9363653	170248	Prob > F
C. Total	59	34589056		<.0001

Expanded Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1419.6	1224.1	-1.16	0.2500
Carts	124.1	33.8	3.67	0.0005
Rain?[No]	650.9	59.7	10.90	<.0001
Rain?[Yes]	-650.9	59.7	-10.90	<.0001
High Temp	21.1	13.2	1.60	0.1161
(Carts-13.9)*Rain?[No]	78.2	33.9	2.31	0.0246
(Carts-13.9)*Rain?[Yes]	-78.2	33.9	-2.31	0.0246

(12) The p-value for the intercept shown in the output implies that, assuming the validity of the standard multiple regression model,

a) The probability that the population intercept is zero is 0.25.
b) The probability that the population intercept is zero is 0.75.
c) The probability that the population intercept is larger than –1420 is 0.25.
d) The probability of an intercept of this magnitude is 0.25 if the population intercept is zero.
e) The probability of an intercept of this magnitude is 0.75 if the population intercept is zero.

(13) If Carts *and* its interaction with Rain were dropped from the model, then

a) Temperature would have a larger coefficient.
b) Temperature would have a smaller coefficient.
c) The R^2 of the model would decrease significantly.
d) The R^2 of the model would be essentially unchanged.
e) Because of possible collinearity, we cannot measure the size of the effect.

(14) The analysis of variance summary shows that

a) This collection of predictors explains significant variation in sales.
b) This collection of predictors does not explain significant variation in sales.
c) Only a subset of these predictors has significant effects on sales.
d) Each predictor has a significant effect on sales.
e) The residuals contain only random variation; the model captures all predictable variation.

(15) Average sales for *one* cart on a rainy day with high temperature 90 degrees is about

a) 50 units.
b) 125 units.
c) 200 units.
d) 250 units.
e) 650 units.

(16) A typical June day has temperature 85 degrees and a typical August day has temperature 95 degrees. If the vendor has 15 carts on duty and it does not rain, then with 97.5% probability the average difference in sales on a typical day in August minus a typical day in June will be

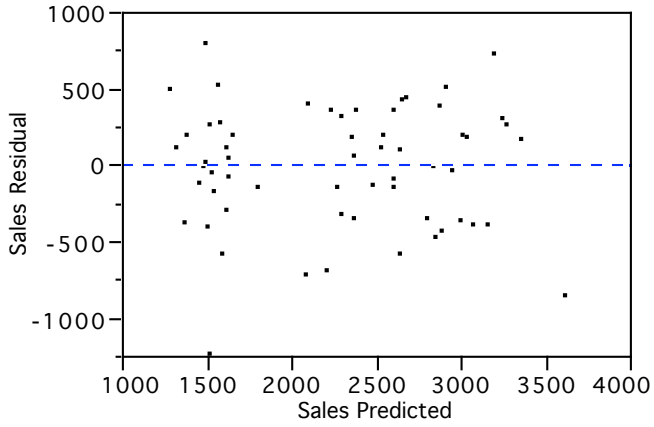
a) Less than 1000 units.
b) Less than 475 units.
c) Less than 211 units.
d) Zero.
e) Negative.

(17) When comparing sales on 85 degree rainy days to sales on 85 degree dry days, with 14 carts on duty, total beverage sales on rainy days *decrease* from sales on dry days on average by about

a) 78 units.
b) 156 units.
c) 650 units.
d) 1300 units.
e) Cannot be answered from the shown output; this analysis requires a two-sample t-test.

(18) The model predicts that average sales tomorrow will be 3000 units. The vendor has on hand 3400 units. The probability that the vendor will run out, using standard assumptions, is about

a) 1/6
b) 2/6
c) 3/6
d) 4/6
e) 5/6



(19) The plot of residuals on the fitted values shown above indicates that

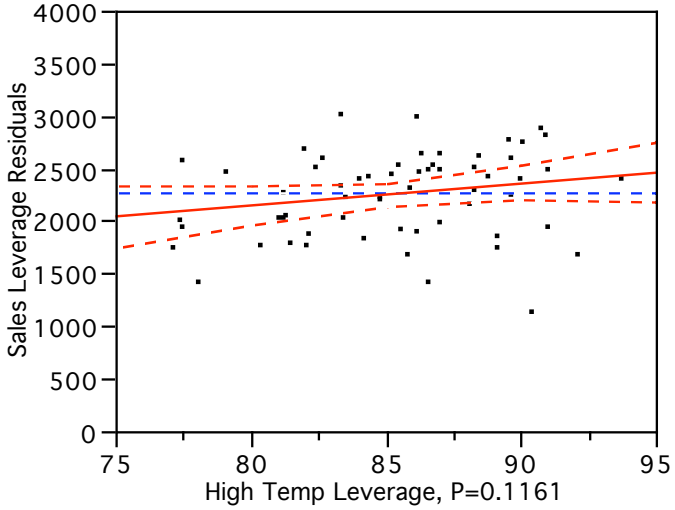
a) An outlier has lowered the accuracy of the fit and should be set aside.
b) An outlier has improved the accuracy of the fit but should be set aside.
c) The data are clustered and dependent.
d) The errors of the model are not normally distributed.
e) The model appears to meet the usual assumptions.

(20) If the vendor wants to learn if the effect of temperature on sales depends on whether the day is rainy or not, he should

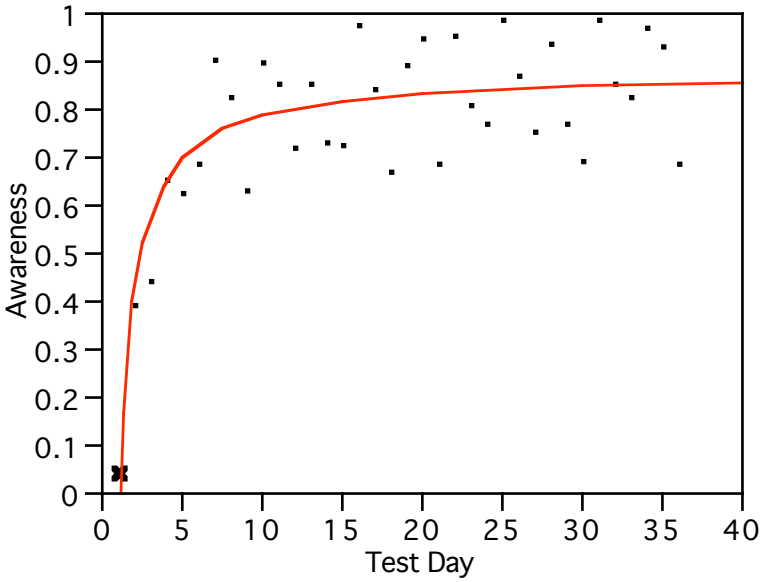
a) Inspect the leverage plot for the variable "Rain?".
b) Plot of the residuals on the categorical predictor "Rain?".
c) Add an interaction between "Rain?" and Temperature.
d) Inspect the profile plot for "Rain?".
e) Not pursue this further since temperature has no significant effect on sales.

(21) The following leverage plot (for temperature) shows that

a) This predictor has a significant impact on the response.
b) The effects of this predictor have been diluted by collinearity.
c) This predictor is not significant because of the impact of outliers.
d) A transformation of this predictor would lead to a significant effect.
e) This predictor has a weak, but insignificant effect on sales.



(Questions 22-27) A business gathered data on the level of customer awareness of a product sold in its outlets. The following analysis considers how the level of awareness (as a proportion of the customers ranging from 0 to 1) of this product changed over the length of time that the product was advertised in its outlets. The predictor *Test Day* is 1 for the day that the product was introduced.



Transformed Fit to Recip

Awareness = 0.877 - 0.90 Recip(Test Day)

Summary of Fit

RSquare	0.713
Root Mean Square Error	0.103
Mean of Response	0.774
Observations	36

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.88	0.02	42.67	<.0001
Recip(Test Day)	-0.90	0.10	-9.00	<.0001

(22) If the program were continued for many more days, the fitted model implies that the level of awareness would top out at

a) 71%
b) 80%
c) 88%
d) 100%
e) Cannot be answered from the shown output.

(23) Management has decided from related cost-benefit calculations that each 1% increase in awareness (e.g., going up from 0.35 to 0.36) is worth $10,000 in subsequent sales. If the cost of running this program is $20,000 per day, then retrospectively the program should have been stopped

a) Around day 3.
b) Around day 7.
c) Around day 11.
d) After day 20.
e) It should never have been run.

(24) On test day 10, the company gathered a sample of 100 customers to check the information supplied by its data-gathering vendor. Based on the fitted model, on average how many of these 100 should be aware of the program.

a) Less than half
b) 79
c) 88
d) 89
e) All

(25) In order to assess the effect of the observation marked with an "x" (the first test day) on the fitted model, we should

a) Examine the leverage plot.
b) Plot the reciprocal of awareness on test day.
c) Plot the awareness on the reciprocal of the test day.
d) Look at a sequence plot of the residuals.
e) Examine the normal quantile plot of the residuals.

(26) In order to investigate whether the data show a significantly higher awareness on certain days-of-the-week (e.g., higher on Monday than on Wednesday, say), we should

a) Add an interaction between a day-of-the-week factor and test day to the model.
b) Do a one-way analysis of variance of awareness by day-of-the-week.
c) Check the Durbin-Watson statistic for autocorrelation.
d) Add a categorical factor indicating the day of the week to the model.
e) Fit three separate regressions, one for each day-of-the-week.

(27) The measured data captured by this sequence of surveys is the proportion that were aware of the new product. Because the response data is a proportion, we should suspect that

a) The data are dependent over time.
b) The data lack constant variation over time.
c) The responses are not normally distributed.
d) The degrees of freedom used to measure the significance of predictors are not meaningful.
e) The model should meet the usual assumptions of regression.

(Questions 28-32) A company has conducted a test of 3 new technologies for wireless networking. It tested the methods in a large city, a dense suburb, a flat farm, and hilly park. Each technology was deployed in all 4 locations. To avoid accidental problems due to unforeseen issues, each combination of technology and location was tested 5 times. For each test, analysts computed a score; scores near zero indicate little signal gets through. Scores at 100 or higher are essentially perfect. An analysis of the resulting 60 measurements follows. Redundant interaction terms for Technology 3 are not shown.

Response Score
Analysis of Variance

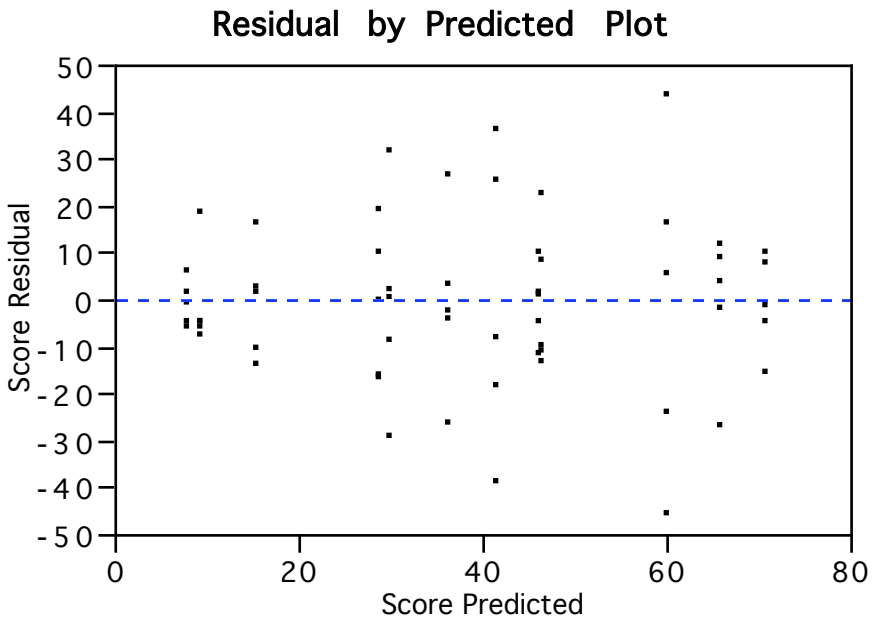
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	11	24474.6	2224.96	6.3151
Error	48	16911.5	352.32	Prob > F
C. Total	59	41386.1		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	37.80	2.42	15.60	<.0001
Technology[1]	5.24	3.43	1.53	0.1326
Technology[2]	4.33	3.43	1.26	0.2128
Technology[3]	-9.52	3.43	-2.79	0.0075
Location[farm]	7.55	4.20	1.80	0.0784
Location[hilly]	-27.36	4.20	-6.52	<.0001
Location[suburban]	13.07	4.20	3.11	0.0031
Location[urban]	6.77	4.20	1.61	0.1144
Technology[1]*Location[farm]	14.83	5.94	2.50	0.0159
Technology[1]*Location[hilly]	-0.71	5.94	-0.12	0.9058
Technology[1]*Location[suburban]	-9.99	5.94	-1.68	0.0987
Technology[1]*Location[urban]	-4.13	5.94	-0.70	0.4983
Technology[2]*Location[farm]	-20.23	5.94	-3.41	0.0013
Technology[2]*Location[hilly]	-5.89	5.94	-0.99	0.3262
Technology[2]*Location[suburban]	15.30	5.94	2.58	0.0131
Technology[2]*Location[urban]	10.80	5.94	1.80	0.0742

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Technology	2	2	2755.8	3.9	0.0267
Location	3	3	15328.3	14.5	<.0001
Technology*Location	6	6	6390.4	3.0	0.0137



Technology*Location, Least Squares Means Table

Level	Least Sq Mean	Std Error
1,farm	65.4	8.4
1,hilly	15.0	8.4
1,suburban	46.1	8.4
1,urban	45.7	8.4
2,farm	29.4	8.4
2,hilly	8.9	8.4
2,suburban	70.5	8.4
2,urban	59.7	8.4
3,farm	41.2	8.4
3,hilly	7.5	8.4
3,suburban	36.0	8.4
3,urban	28.3	8.4

(28) The estimated signal strength for Technology 1 in a suburban area is about
a) –9.99
b) 5.2
c) 33
d) 43
e) 46

(29) Does the performance of Technology 2 depend on the location?
a) Yes, the Effect Test for the interaction is significant.
b) No, the Effect Test for the interaction is not significant.
c) Yes, the location effect is significant.
d) No, the technologies are all about the same
e) Yes, some p-values for Technology[2]*Location are less than 0.05/3.

(30) Regarding Technology 1, this analysis indicates that Technology 1
a) Is not significantly different from the other two.
b) Is significantly better than either of the others.
c) Works well in the open farm location, but is not the best otherwise.
d) Produces an average signal score of 43 on this specific scale.
e) Because of an interaction, we cannot interpret these results for Technology 1.

(31) An important conclusion of this analysis is that
a) The choice of the best technology depends on the location.
b) Significant differences exist among the technologies, with Technology 1 the best.
c) No significant difference exists among the 3 technologies at any location.
d) None of the technologies obtains a score above 100 on any occasion.
e) The largest difference among the methods is found in the hilly location.

(32) It was later learned that all of the measurements in the farm location were obtained during a rainy period. All of the other measurements were obtained under dry conditions. Rain is known to reduce the test score. Because of this, we should conclude that
a) The measured significance values are wrong because of the resulting dependence.
b) The full experiment needs to be repeated to insure consistent weather conditions.
c) The rainy conditions have exaggerated the differences among the technologies.
d) The results for the farm location may be due as much to weather as location.
e) A variable indicating the presence of rainfall should be added to the model.

(Questions 33-42) The marketing research group at a large pharmaceutical firm conducted an internal study of the impact of its own marketing program. This particular program uses so-called "detail representatives" who call on doctors to promote the prescription of certain medications. When a detail rep visits a doctor (each visit is called a "detail"), they may also leave samples. In this example, the medication of interest is *Nosorr*, a product designed to reduce inflammation and the pain associated with arthritis. For each of 15 communities, the firm tracked the use of its products for a sample of about 19 doctors at each location. The total sample has 282 observations. For every sampled doctor, the firm has the following information for the most recent quarter (3 months) of activity:

Rx	the number of prescriptions for *Nosorr* written by the MD during the quarter
SMSA	Name of the community (Standard Metro Sampling Area)
Est visits	Estimated number of visits that might be treated with this medication
Region	Abbreviated name for sales district (**E**ast, **W**est, **R**ocky Mtn, and **C**entral)
Age	Of the MD in years.
Specialty	Of the MD: GP (general practice), RH (rheumatologist), PD (pediatrician)
Details	The number of visits by *Nosorr* detail representatives during the quarter
Samples	The number of samples of *Nosorr* left by detail representatives

A preliminary analysis of the response produced the following correlations and least squares regression fits. The response is the (natural) log of the number of prescriptions written for *Nosorr*. All logs in this analysis are natural (base e) logs.

Correlations

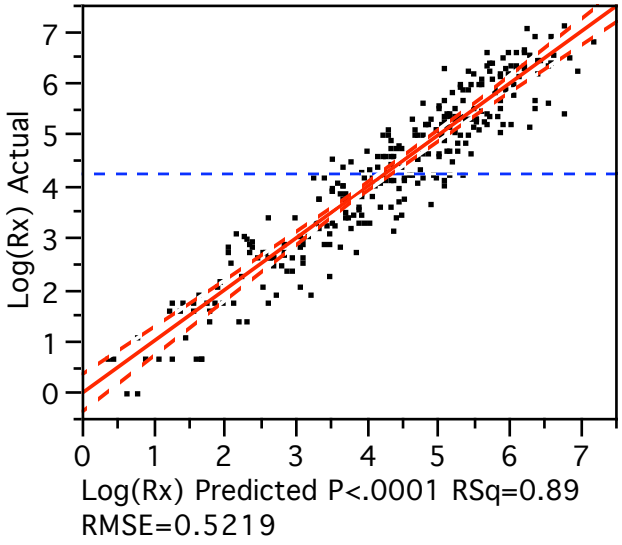
	Log(Rx)	Log(Details)	Log(Samples)	Log(Visits)	Age
Log(Rx)	1.00	0.71	0.75	0.88	0.01
Log(Details)	0.71	1.00	0.95	0.63	0.01
Log(Samples)	0.75	0.95	1.00	0.60	0.01
Log(Visits)	0.88	0.63	0.60	1.00	0.00
Age	0.01	0.01	0.01	0.00	1.00

Least squares fits:

$$\text{Log(Rx)} = 2.47 + 1.96\ \text{Log(Details)}$$

$$\text{Log(Rx)} = -3.03 + 1.70\ \text{Log(Samples)}$$

The firm also estimated the following multiple regression. The response in this multiple regression is also Log(Rx).



Effect Tests

Source	DF	Sum of Squares	F Ratio	Prob > F
Log(Details)	1	5.807	21.32	<.0001
Log(Samples)	1	0.029	0.10	0.7406
Log(Visits)	1	77.183	283.39	<.0001
Region	3	5.925	7.25	0.0001
Age	1	0.161	0.59	0.4417
Specialty	2	32.659	59.95	<.0001

Expanded Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.620	0.565	-1.10	0.2739
Log(Details)	0.773	0.167	4.62	<.0001
Log(Samples)	-0.050	0.150	-0.33	0.7406
Log(Visits)	1.044	0.062	16.83	<.0001
Region[C]	-0.074	0.049	-1.50	0.1338
Region[E]	0.246	0.054	4.53	<.0001
Region[R]	-0.059	0.070	-0.83	0.4059
Region[W]	-0.114	0.060	-1.89	0.0602
Age	-0.002	0.003	-0.77	0.4417
Specialty[GP]	0.097	0.044	2.19	0.0292
Specialty[PD]	-0.632	0.060	-10.54	<.0001
Specialty[RH]	0.535	0.056	9.57	<.0001

(33) The F-ratio for the regression of Log(Rx) on Log(Samples) is
a) 0.10
b) A value less than 4 (but one that cannot be determined precisely)
c) 280
d) 360
e) The square of the t-statistic for the slope and thus cannot be determined from the output.

(34) Management has decided to institute a 10% increase nationwide in the number of samples distributed by its detail force during their usual rounds. The shown output indicates that the effect on the number of prescriptions produced by this increase in the number of samples is expected to be
a) Significantly negative.
b) Zero.
c) Less than 2.5%.
d) About 17%, but one cannot determine the significance from this output.
e) About 17%, and statistically significant.

(35) Based on this analysis, which specialty is most responsive to detailing?
a) GP
b) RH.
c) PD.
d) There are no significant differences among the specialties in this regard.
e) This analysis cannot answer this question.

(36) If the level and distribution of promotion is maintained then a GP physician in the East region whose practice grows from 150 visits to 165 visits would be expected to write
a) About the same number of prescriptions for *Nosorr*.
b) About 15 more prescriptions.
c) 10% more prescriptions for *Nosorr*.
d) 13% more prescriptions for *Nosorr*.
e) Cannot be determined from the shown output.

(37) The estimated coefficient for the log of the number of details in the multiple regression implies that, on average and adjusting for the other factors in the model,
a) Each detail leads to about 0.77 more prescriptions.
b) Each detail leads to about a 0.77% increase in prescriptions.
c) Each 1% increase in details leads to about a 0.77% increase in prescriptions.
d) There is less than a 0.0001 chance that detailing is effective in increasing sales.
e) The effect is not significantly different from zero and ought not be interpreted further.

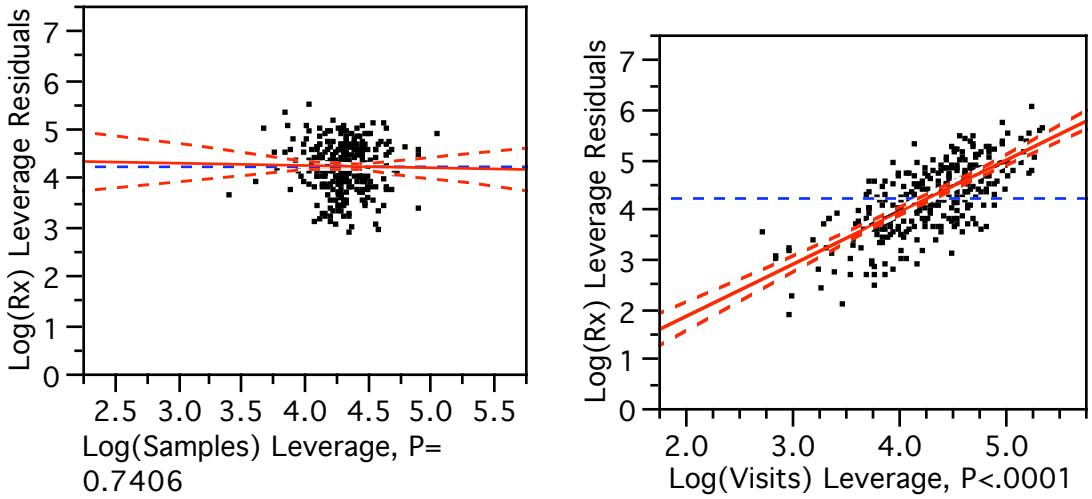
(38) To examine the statistical significance of differences between regions, we should
a) Analyze the 4 groups of data separately in a one-way analysis of variance.
b) Examine the associated Tukey-Kramer comparison of these 4 coefficients.
c) Add the variance inflation factors to reveal the extent of collinearity.
d) Use the Effect Test summary rather than any of the reported t-statistics.
e) Treat these values like any regression coefficient and use the shown t-statistics.

(39) A competitive model developed by another marketing research group at the firm modeled the number of scripts directly. This model (not shown) obtained an R^2 of 81%. Direct comparison of the R^2 of that model fit to the R^2 of the shown multiple regression

a) Indicates that the other model fits significantly worse than the shown model.
b) Indicates that the other model fits about the same as the shown model.
c) Indicates that the other model fits significantly better than the shown model.
d) Indicates that the other model is not properly specified (i.e., uses the wrong predictors).
e) Is inappropriate.

(40) The following pair of leverage plots indicates that

a) Both log(samples) and log(visits) are statistically significant predictors.
b) Leveraged outliers have distorted the slope for log(visits).
c) Leveraged outliers have distorted the slope for log(samples).
d) Collinearity has a stronger effect on log(samples) than log(visits).
e) The data are not normally distributed; too many values lie outside the bounds.



(41) Management felt that the detailing force in one region was more effective in generating scripts per detail than the others. To resolve this issue the analyst should

a) Add the interaction of log(details) and Region to the fitted model.
b) Run a one-way analysis of variance of the level of detailing by Region.
c) Remove the sampling factor from the model to clarify the effect of detailing.
d) Run 4 separate regressions of log(Rx) on log(details), one for each sales district.
e) Note that the Effect Test for Region is significant, indicating large differences.

(42) Which of the following would **not** be an appropriate next step for the shown multiple regression analysis (i.e., which would be a foolish next step)

a) Compare boxplots of the residuals, grouped by region.
b) Remove the predictor *Age* from the regression analysis.
c) Remove the predictor Region[R] from the regression analysis.
d) Examine a normal quantile plot of the residuals.
e) Compute variance inflation factors to quantify the effects of collinearity.