Statistics Waiver Exam

August 20, 2012

Bring your Penn student ID, #2 pencils, eraser, and calculator to the exam. You may also bring *one page of handwritten notes* (8.5"´11" or A4, using both sides as you like) to the exam.

When you receive an answer sheet *before* the exam begins …

Fill in your name and Penn student ID number.
Your ID number appears in bold on your ID.

Mark the "bubbles" under the letters of your name *and* student ID number.
Failure to do so will lead to a score of zero.

Use a **#2 pencil**. Erase any changes completely.

Turn off your phone.
You are not allowed to make or receive a call during the exam. Use of your phone (for calls or messages) during the exam is grounds for dismissal from the exam.

Once the exam begins …

Choose the **one *best* answer** for each question. The exam has **50 questions**.
Picking more than one answer is scored as an error.

You may consult **1 page of handwritten notes** during the exam.
No other reference materials are permitted.

You may use a basic **calculator** or graphing calculator.
No laptops, phones, or computers are allowed.

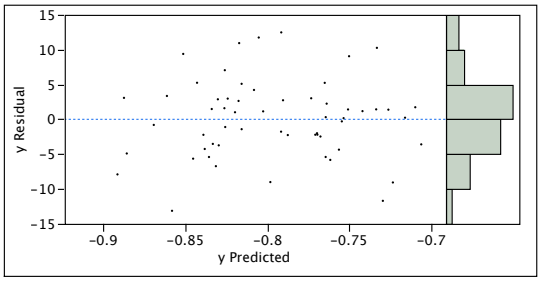
You have **two hours** for the exam. The **computer output** associated with one or more items should be considered an essential part of the questions. Throughout, the word "significant" implies "statistically significant". The abbreviation SRM stands for standard definition of a 'simple regression model'; MRM for 'multiple regression model.' All logs are natural logs (logs using base *e*).

Your **score** is the number of correct answers. The questions are equally weighted. Some questions may be dropped and not counted as part of the overall score. There is no deduction for incorrect answers. Regardless of what you write on your copy of the exam, only the marked answers on the grade form will be considered.

STOP

*Do **not** turn the page until you are instructed*.

1. A least squares regression produced the residual plot shown at the right. The 95% prediction intervals for new observations from this model have the form of a predicted value
   a) ±0.1
   b) ±1
   c) ±5
   d) ±15
   e) ±10



2. The plot that is *most useful* for identifying the presence of unequal error variation in the residuals of a multiple regression model is the
   a) Scatterplot of the residuals versus the response y.
   b) Scatterplot of residuals on fitted values (a.k.a., predicted values).
   c) Normal quantile plot of the residuals.
   d) Histogram of y.
   e) Sequence plot of the residuals in time order.

3. To check whether data used in a simple regression meet the assumption of normality implied by the SRM, one should inspect the
   a) Normal quantile plot of the residuals.
   b) Normal quantile plot of the response y.
   c) Normal quantile plot of the explanatory variable x.
   d) Scatterplot matrix of y and x.
   e) Boxplots of y and x.

4. In a simple regression of salary on a dummy variable indicating the sex of the employee (coded as 1 for male and 0 for female), the coefficient of the dummy variable estimates the
   a) Average salary of male employees.
   b) Standard deviation of the salary of male employees.
   c) Average salary of female employees.
   d) Difference between average salaries of male and female employees.
   e) Ratio of the average salary of male employees to female employees.

5. The linear regression of Y on X produces the fitted model Y = 2.33 - 2.55 X based on $n = 85$ observations. Assume that the Simple Regression Model holds. To test the null hypothesis that the underlying intercept β_0 is equal to zero we need to know the
   a) Standard deviation of the residuals.
   b) Mean of the response Y.
   c) Standard deviation of the response Y.
   d) Standard error of the intercept.
   e) Standard deviation of X.

6. If the p-value of the slope in a multiple regression is 0.27, then we should conclude that
   a) The estimated slope is within 2 standard errors of zero.
   b) The estimated slope is an extrapolation from the rest of the data.
   c) The population slope β_1 is zero.
   d) We should reject the null hypothesis H_0: $\beta_1 = 0$.
   e) Increasing the sample size would increase the p-value for the slope.

7. The presence of heteroscedasticity in a multiple regression implies that
   a) The R^2 statistic is close to 0.
   b) The R^2 statistic is close to 1.
   c) The overall F-ratio of the model is close to 4.
   d) The model predicts observations at some x's more precisely than at others.
   e) The explanatory variables in the model are correlated with each other.

8. An observation of the explanatory variable x is most capable of influencing the slope of a simple least squares regression when it is
   a) Negative.
   b) Exactly in the center of the data in the x-direction.
   c) Large and positive.
   d) Close to other data points.
   e) Much smaller or much larger than the other values of x.

9. Which of the following statements does **not** follow from the underlying assumptions made by the simple regression model?
   a) The confidence interval for β_1 is b_1 plus or minus twice its standard error.
   b) Approximately 68% of the data lie within one SD(errors) of the regression line.
   c) Approximately half of the data lie beneath the least squares line.
   d) A residual plot will show no structure or pattern.
   e) The explanatory variable x is normally distributed around its mean.

10. The fitted regression equation of log(y) against log(x) is log(y) = 2.3 + 0.1 log(x), with residual standard deviation = 0.1 on the log(y) scale. (Recall that all logs are natural logs, logs to base e.) Assuming the SRM holds true and we are not extrapolating, the approximate 95% prediction interval for y when x = 6 is about
    a) (2.31, 2.64).
    b) (2.28, 2.68).
    c) (9.77, 14.57).
    d) (10.12, 14.06).
    e) (10.11, 13.75).

11. If the overall F-test found in the analysis of variance summary for a multiple regression is significant, then
    a) At least one of the t-statistics for a slope must also be significant.
    b) All of the t-statistics for the slopes must be significant.
    c) We can reject H_0 that all of the slopes are simultaneously zero.
    d) The regression must be expanded to include at least one interaction.
    e) There is no chance of having made a Type I error.

12. Assuming the multiple regression model holds, if the 95% confidence interval for the intercept β_0 is the interval (-27, 9), then the
    a) Intercept β_0 is zero.
    b) t-statistic for the intercept is between -2 and +2.
    c) Intercept β_0 is greater than zero.
    d) Residual standard deviation is about 9.
    e) Standard error for the intercept is about 18.

13. Which of the following regressions provides a global (rather than at a point) estimate of the elasticity of changes in y with respect to x?
    a) Regression of log y on x.
    b) Regression of y on log x.
    c) Regression of log y on log x.
    d) Regression of $1/y$ on x.
    e) Regression of y on x.

14. In a simple regression with response y measured as average cost and one predictor x that is the reciprocal of the number produced, the intercept estimates
    a) Marginal costs.
    b) Shutdown costs.
    c) Fixed costs.
    d) Rate of growth in the response.
    e) Total cost, adjusted for run size.

15. The presence of an interaction in a multiple regression implies that
    a) Several variables in the model are collinear.
    b) The slope for one explanatory variable depends on the value of another.
    c) The estimated model suffers from autocorrelation.
    d) The intercept for one group differs from the intercept for another.
    e) The fitted model conceals the presence of a confounding variable.

16. By choosing values of the predictor in a simple regression model that are more spread out rather than concentrated along the x-axis, one
    a) Decreases the standard deviation of the residuals in the fitted model.
    b) Decreases the r^2 of the fitted model.
    c) Violates the assumption of equal variance.
    d) Violates the assumption of normality of the errors.
    e) Obtains a more precise estimate of the underlying population slope.

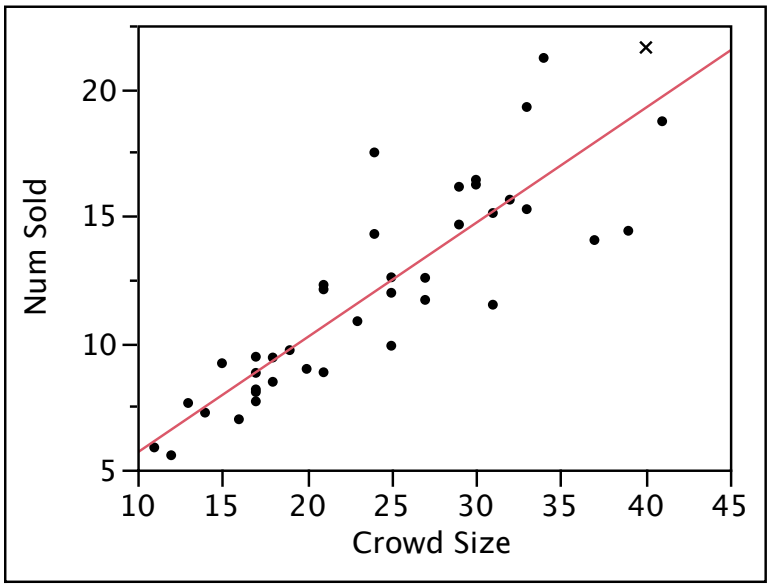
17. Collinearity affects a regression by
    a) Decreasing the standard error of estimated slopes.
    b) Increasing the magnitude of t-statistics (t-ratios) for slopes.
    c) Increasing the length of confidence intervals for slopes
    d) Decreasing in the residual standard deviation.
    e) Increasing the residual standard deviation.

18. The same simple regression model was fit to two different samples drawn from the *same* population. One of the two samples had 250 cases; the other had 50 cases. We can be quite certain that the regression fit to the larger sample will have
    a) Smaller residual std. deviation than the regression fit to the smaller sample.
    b) Larger r^2 than the regression fit to the smaller sample.
    c) Shorter 95% CI for the slope than the regression fit to the smaller sample.
    d) Residuals that are more likely to be normally distributed.
    e) All of the above.

19. In a regression with a categorical variable, comparison boxplots of the residuals grouped according to the categorical variable are most useful to detect
    a) The presence of a confounding factor.
    b) A lack of normality.
    c) A statistically significant difference among the groups.
    d) The presence of unequal residual variance.
    e) The absence of an important explanatory variable.

20. In a time series regression, dependence among the underlying model errors is most likely indicated by inspecting
    a) Leverage plots.
    b) The scatterplot of the residuals on the fitted (predicted) values.
    c) A normal quantile plot of the residuals.
    d) Comparison boxplots of the residuals when grouped by size.
    e) The scatterplot of the residuals on the lag of the residuals.

(Q21-30) A food vendor would like to anticipate the consumption of beverages at a baseball stadium. If the vendor prepares too few beverages, it loses the opportunity to sell more and causes long lines at service counters. If it prepares too many beverages, the excess that is not sold is lost. The following data show the number of cups of soft drinks sold (*Num Sold*, in thousands) and the official crowd size *Crowd Size* (also in thousands) at 40 recent games. For example, a value of 30 for *Crowd Size* indicates that 30,000 customers were at the game. For each game, the vendor had prepared an excess of beverages so that there were no constraints on the customer demand.



RSquare	0.766336
Root Mean Square Error	2.052372
Mean of Response	12.14365

Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.1429295	1.037463	1.10	0.2775
Crowd Size	0.4522392	0.04051	11.16	<.0001*

21. The fitted model implies that
    a) 76 percent of the observations are close to the fitted regression line.
    b) The standard deviation of beverage sales is about 2 cups sold per person.
    c) An empty stadium – a game that no one attended – generates no sales.
    d) The vendor on average sells about 1 cup for every 2 customers at the game.
    e) None of the above.

22. If 30,000 customers are expected to attend a game, then the fitted model implies that the vendor will sell on average about
    a) 12,140 cups.
    b) 13,600 cups.
    c) 14,700 cups.
    d) 16,900 cups.
    e) 19,000 cups.

23. Suppose that the fitted model predicts sales of 16,000 for a coming game based on the anticipated attendance. Assuming that this number of customers do attend and this model and the assumptions of the SRM hold, how many cups of beverage should the vendor prepare in order to meet demand with 97.5% probability
    a) 16,000 cups.
    b) 17,000 cups.
    c) 18,000 cups.
    d) 19,000 cups.
    e) 20,000 cups.

24. A special promotion is expected to draw 5,000 more customers to a coming game. Based on the fitted model and the assumptions of the SRM, the vendor can expect the presence of these 5,000 additional customers to produce how many more sales, with 95% confidence, on average?
    a) From 700 to 1560 more cups.
    b) From 1,850 to 2,670 more cups.
    c) From 2,400 to 4,400 more cups.
    d) From 1,400 to 5,400 more cups.
    e) The promotion will not significantly affect the demand for beverages.

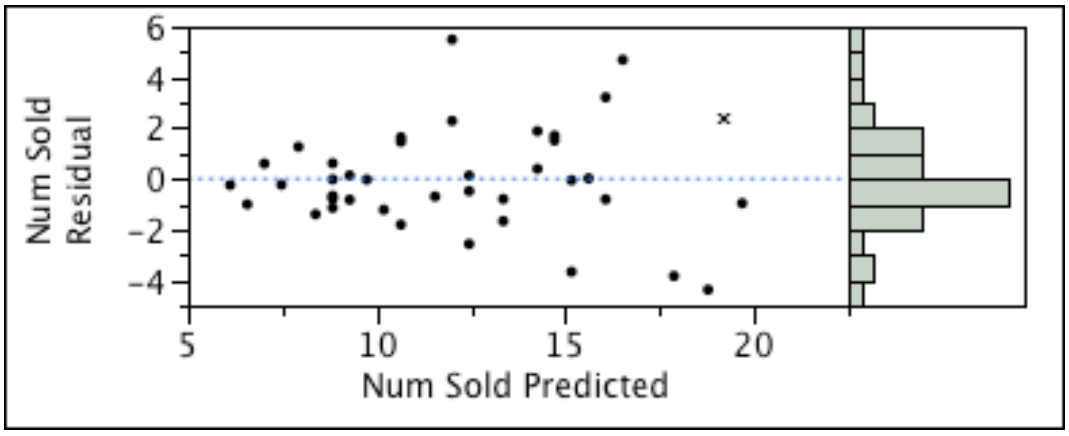
25. If this analysis were repeated, but with the data recorded as the number sold and attendance as actual counts rather than expressed in thousands, then which of the following features of the model would change?
    a) R^2
    b) Residual standard deviation
    c) Estimated slope
    d) t-ratio for the slope
    e) None of these features would change.

26. This vendor operates concessions at many stadiums. If the vendor assumes that the relationship between crowd size and sales of beverages is the same at all these stadiums, it can use a much sample of observations to estimate this regression (say, 200 rather than 40 observations). If the vendor is correct in this assumption, then we can be assured that, compared to the model with $n = 40$ observations, that the
    a) Standard deviation of the residuals when fit to 200 observations will be larger.

b) r^2 when fit to 200 observations will be smaller.
c) Slope when fit to 200 observations will be larger than the slope in this model.
d) 95% confidence interval for the slope when fit to 200 observations is more likely to contain the true slope than the 95% interval from this model.
e) 95% confidence interval for the slope when fit to 200 observations will be shorter than the interval for this model.

27. If the game marked as an "×" in the figure is removed from the fitted model, then *we can be sure* that
    a) The standard deviation of the residuals will increase.
    b) The estimated slope will increase.
    c) The estimated slope will decrease
    d) The estimated intercept will decrease.
    e) The standard error of the slope will decrease.

28. The vendor observed that the largest crowds came to games on the warmest days. That is, the crowd size increases with the temperature at the time of the game. It is also known that beverage consumption increases on warmer days. These features of the data suggest that if temperature is added to the previous fitted model
    a) The standard error of *Crowd Size* will decrease.
    b) The estimated slope for *Crowd Size* will decrease.
    c) The estimated slope for *Crowd Size* will increase.
    d) The estimated slope for *Crowd Size* will remain unchanged.
    e) The r^2 of the model will increase significantly.

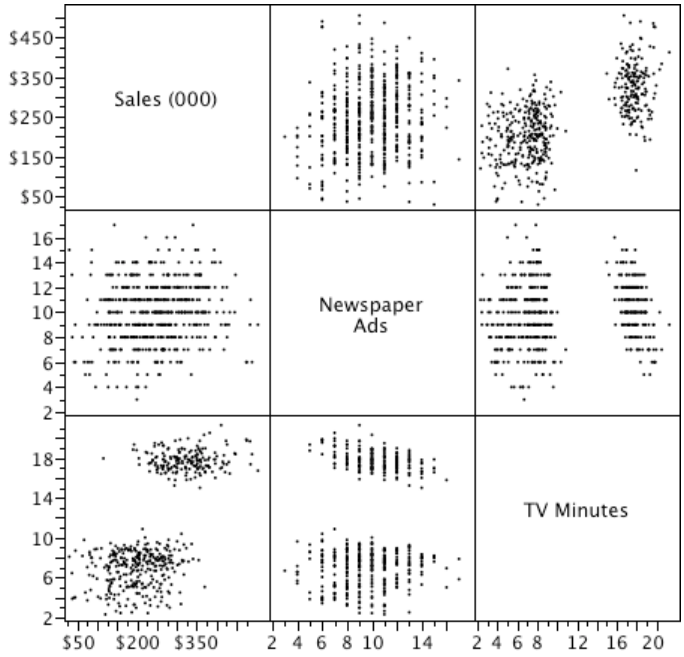


29. The diagnostic plot for this simple regression shown immediately above indicates that
    a) The model meets the assumed conditions of the SRM.
    b) The sample size is too small for such plots to be useful.
    c) The model omits a variable that would lead to more accurate predictions.
    d) Predictions from the model are more accurate than claimed for small crowds.
    e) The data are approximately normally distributed.

30. About half of these 40 games are day games, and the others are night games. If the vendor feels that the number of drinks sold per customer depends on whether the game was played during the day versus at night, then the vendor should
    a) Add a day/night categorical variable to the model.

b) Add a day/night categorical variable and its interaction with *Crowd Size* to the model.
c) Fit a regression of *NumSold/CrowdSize* on *1/CrowdSize*.
d) Use a log/log model to capture the differential elasticities of demand.
e) Leave the model as is since adding categorical terms will introduce collinearity.

(Q31-38) The marketing division of a consumer products company decided to assess the success of its promotions. To this end, it recorded the information in 480 communities spread over 4 geographical regions (East, North, South and West). The variables include the number of advertisements in local newspapers (*Newspaper Ads*) and the number of minutes of television ads shown in each community (*TV Minutes*). Two types of TV commercial, labeled "A" and "B", were shown and are identified by the categorical variable *Commercial*. The response in the shown model is the level of sales in a community, measured in thousands of dollars (for example, *Sales* = 30 implies sales of $30,000). Tables in the output summarize a multiple regression model.



R^2	0.49164
Residual Std Dev (s_e)	63.9206

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	1869042.7	311507	76.2406
Error	473	1932603.7	4086	**Prob > F**
C. Total	479	3801646.4		<.0001*

Effect Tests (Partial F tests)

Source	DF	Sum of Squares	F Ratio	Prob > F
Newspaper Ads	1	29197.7	7.1461	0.0078*
TV Minutes	1	1146332.5	280.5621	<.0001*
Commercial	1	73162.6	17.9064	<.0001*
Region	3	31037.8	2.5321	0.0564

Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	63.68000	20.07871	3.17	0.0016*
Newspaper Ads	3.87700	1.45032	2.67	0.0078*
TV Minutes	11.40641	0.68098	16.75	<.0001*

Term	Estimate	Std Error	t Ratio	Prob>|t|
Commercial[A]	25.23405	5.96325	4.23	<.0001*
Region[EAST]	-4.01966	8.25747	-0.49	0.6266
Region[NORTH]	7.46986	9.15808	0.82	0.4151
Region[SOUTH]	25.14592	10.85476	2.32	0.0210*

31. The scatterplot matrix shown on the previous page indicates that
    a) A regression analysis using these variables suffers from severe collinearity.
    b) *TV Minutes* is typically less than 10 or near 20.
    c) Numerous leverage points distort the relationship between *Sales* and *TV Minutes*.
    d) The two types of commercials define two groups of observations in the data.
    e) *Sales* are negatively correlated with the number of newspaper ads.

32. It was claimed that each minute of TV advertising adds $10,000 to local sales, on average, holding other relevant factors constant. The summary of this regression implies that the effect of each additional minute
    a) Is significantly less than $10,000.
    b) Is not significantly different from $10,000.
    c) Is significantly more than $10,000.
    d) Depends on the location of the advertisement.
    e) Depends on which commercial is shown.

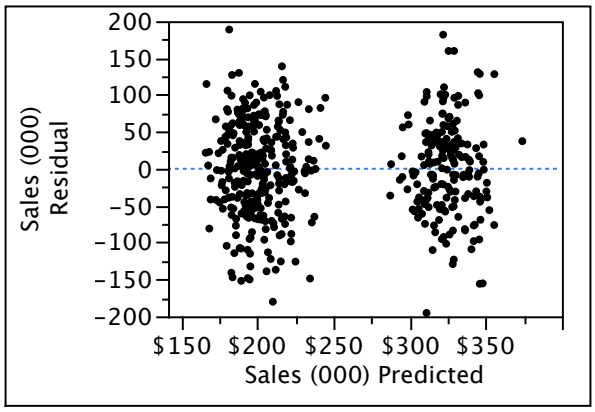
33. Given a consistent level and type of newspaper and television advertising, has the promotion been significantly more successful in some regions than others?
    a) Yes, sales are significantly higher in the South than in other regions.
    b) Yes, both estimates for *Commercial* are significant.
    c) No, because the effect for the North is too close to zero.
    d) No, because the effect test for *Region* is not significant.
    e) We cannot tell without the marginal analysis of variance of *Sales* by *Region*.

34. If the type of commercial in a community is changed from "A" to "B" while maintaining the same level of newspaper and TV advertising, then this model predicts sales would on average
    a) Increase $12,620.
    b) Increase $25,240.
    c) Decrease $12,620.
    d) Decrease $25,240.
    e) Increase $38,534.

35. Two communities experience the same number of newpaper ads and have the same amount and type of TV advertising. One of these communities is in the East and the other is in the West. If the MRM holds for these data, then with 95% confidence, average sales
    a) Are higher in the West than East by between –$20,500 to $12,500.
    b) Are higher in the West than East by between $23,500 to $103,800.
    c) Are higher in the West than East by between –$12,300 to $4,200.
    d) Are higher in the East than West by between –$12,500 to $20,500.
    e) Are higher in the West than East by between –$131,900 to $123,800.

36. An analyst noted that two communities in the South both had 10 newspaper ads and 4.5 minutes of Commercial "A". Sales in the one community were about $210,000, whereas those in the other were $270,000. We should conclude from this insight that
    a) The amount of spending on advertising should be increased in both communities.

b) There is evidence of cannibalization of sales from one community to another.
c) The sales manager in the second community deserves a bonus.
d) Commercial A was significantly more effective in the second community.
e) These two levels of sales are comparable.



37. The plot of the residuals of the fitted model on the predicted values shown immediately above implies that
    a) These data do not conform to the assumptions of multiple regression.
    b) The data lack constant variance for the two types of commercials.
    c) The predicted values from the model fall into two distinct ranges of values.
    d) The distribution of the error terms underlying the model is not normal.
    e) The model suffers from substantial collinearity, leading to vertical clusters.

38. Which of the following actions would **not** be an appropriate next step for the analysis of this data? (i.e., Which of the following actions would be inappropriate?)
    a) Add the interaction between *Commercial* and *TV Minutes*.
    b) Color-code the scatterplots and residual plots by *Location* or *Commercial*.
    c) Check for constant variance in the residuals by grouping by *Location*.
    d) Quantify possible effects of collinearity by adding VIFs to the model summary.
    e) Any of the above actions is appropriate as a next step for this analysis.

(Q39-44) A seller of used automobiles purchases classified advertisements that appear on two web site. The seller like to know if one site generates more sales than the other, and it would also like to know how much revenue these advertisements generate. The seller has collected information over the last 30 weeks. For each week, the seller recorded the number of "web pages" of advertisements shown on each web site. The seller had its salesmen asked potential customers where they learned that a particular car was for sale. The data for this analysis identifies the web site (*Site* is either the "G" or the "B"), the number of web pages shown (*Pages*), and as a response, the number of contacts generated (*Contacts*).

R^2	0.842824
Root Mean Square Error	6.21078

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	11583.301	3861.10	100.0965
Error	56	2160.132	38.57	**Prob > F**
C. Total	59	13743.433		<.0001*

Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5.05992	5.181187	0.98	0.3330
Site[B]	-13.85284	9.996309	-1.39	0.1713
Pages	1.73396	0.357706	4.85	<.0001*
Pages*Site[B]	3.66715	0.817128	4.49	<.0001*

39. The fitted model indicates that, on average, 5 pages of advertising on site *B* will generate on average about
    a) 16 contacts.
    b) 17 contacts.
    c) 18 contacts.
    d) 33 contacts.
    e) None of the above.

40. Based on the fitted model as summarized, advertisements shown on web site B, per added page, generate about
    a) The same number of contacts as those shown on site G.
    b) 28 more contacts than those shown on site G.
    c) 14 more contacts that those shown on site G.
    d) 1.7 more contacts than those shown on site G.
    e) 3.7 more contacts than those shown on site G.

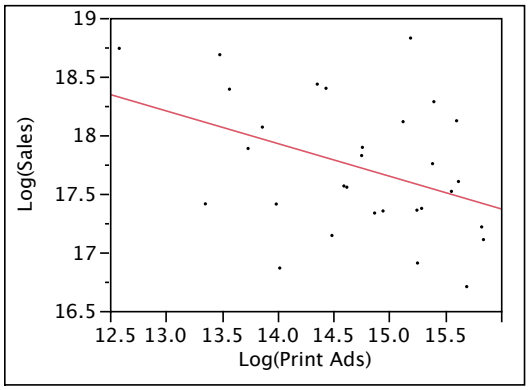
41. The fitted model implies that an increase of 10 more pages of advertising on site B will generate on average about
    a) 18 more contacts.
    b) 36 more contacts.
    c) 54 more contacts.
    d) 170 more contacts.
    e) 0 more contacts.

42. The seller plans to show 13 pages of advertising on each of these sites. This model suggests that the *difference*, on average, in contacts generated by the two sites will be
    a) Not statistically significantly different from zero.
    b) 17 more contacts generated by ads on site B.
    c) 33 more contacts generated by ads on site B.
    d) 17 more contacts generated by ads on site G.
    e) This comparison cannot be made in the presence of an interaction.

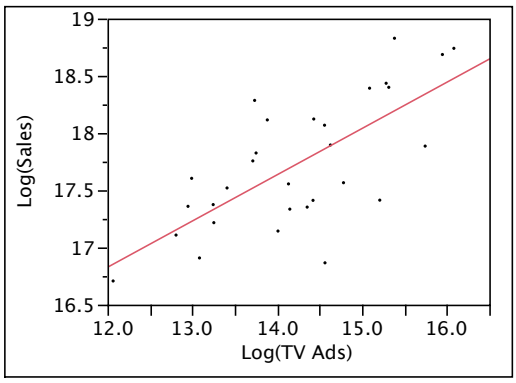
43. An assumption of the multiple regression model is the equality of the variance of the errors in the groups defined by the two sites. To check this assumption, one should
    a) Inspect the normal quantile plot of the residuals.
    b) Run a two-sample t-test of the average of the residuals grouped by site.
    c) Add an interaction term to the fitted model.
    d) Inspect comparison boxplots of the two groups of residuals defined by the sites.
    e) Inspect comparison boxplots of the number of contacts defined by the two sites.

44. Because the data measure the response of the two sites during the same weeks, we should suspect which of the following assumptions of the MRM is false?
    a) The observations are independent.
    b) The effect of each explanatory variable is linear.
    c) The effect of each explanatory variable is constant.
    d) The observations have equal error variance.
    e) The underlying model errors are normally distributed.

(Q45-50) [1999 621] A seller of children's toys has been studying the allocation of its historical advertising between printed advertisements, such as those mailed directly to consumers and in newspapers, and television advertising, particularly concentrated during Saturday morning cartoons. The data give the annual sales of its products and the spending on television and printed advertisements over the past 30 years. All amounts are expressed on the scale of the natural log of US dollars. (The natural log is the logarithm to base e). The output shows two simple regressions and a multiple regression.



Log(Sales) = 21.8 – 0.28 Log(Print Ads)
Standard error for slope = 0.12
SD(residuals) = 0.50 R^2=0.19



Log(Sales) = 12.0 + 0.40 Log(TV Ads)
Standard error for slope = 0.08
SD(residuals) = 0.40 R^2=0.49

Mult Regression, Response Log(Sales)

RSquare	0.58
SD(residuals)	0.37

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	3.89	3.69	1.05	0.3020	0.00
Log(Print Ads)	0.33	0.15	2.28	0.0309	2.89
Log(TV Ads)	0.63	0.12	5.17	<.0001	2.89

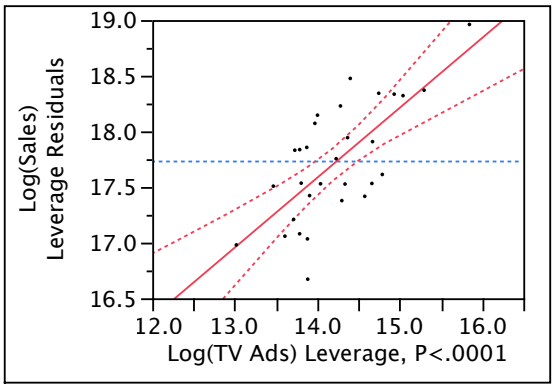
45. The estimated slope in the regression of Log(Sales) on Log(TV Ads) implies that, on average,
    a) Changes in spending for television ads have no significant effect upon sales.
    b) For each additional dollar spent on television ads, sales increase by \$0.40.
    c) For each additional dollar spent on television ads, sales increase by 0.40%.
    d) For each 1% increase in dollars spent on television ads, sales increase by 0.40%.
    e) For each 1% increase in dollars spent on television ads, sales increase by \$40.

46. Given the standard assumptions for the regression of Log(Sales) on Log(TV Ads), if the seller spends \$1.2 million on television advertising, then the probability of its sales being less than \$20 million is
    a) Approximately 2.5%.
    b) Approximately 5%.
    c) Approximately 16%.
    d) Approximately 33%.
    e) More than 50%.

47. Does the estimated multiple regression explain statistically significantly more variation in the log of sales than the model which uses only the log of the television advertising? (Assume that these models meet the required assumptions of regression.)
    a) No, because the change in R^2 is too small to be useful.
    b) No, because the *t*-statistic for the slope is smaller in the multiple regression.
    c) No, the SD of the residuals is not much smaller than using Log(TV Adv) alone.
    d) Yes, Log(Print Ads) adds a significant improvement.
    e) Cannot tell without further output.

48. From the shown regression output indicates that if the seller retains the current level of advertising and increases the amount of printed advertising next year that it can expect (assume the conditions of regression modeling are satisfied)
    a) Sales to fall by a statistically significant amount.
    b) Sales to fall, but not by a statistically significant amount.
    c) Sales to remain at the current level.
    d) Sales to increase, but not by a statistically significant amount.
    e) Sales to increase, by a statistically significant amount.

49. The partial regression leverage plot from the multiple regression shown immediately below and to the right indicates that
    a) Several leveraged observations distort the slope of Log(TV ads).
    b) Log(TV ads) makes a significant contribution to the model.
    c) Log(TV ads) does not make an significant contribution to the model.
    d) The data are contaminated by outliers because many points lie outside the dashed bands.
    e) The fitted model omits time trends that would improve prediction.



50. The plot of the residuals from the multiple regression shown immediately below and to the right indicates
    a) Larger samples are needed in order to find important effects.
    b) The data are autocorrelated.
    c) The data are heteroscedastic.
    d) The data are not normally distributed.
    e) No deviation from the usual assumptions.