August 2009

Solutions to the Statistics 621 Waiver Exam

1. Mark the answer to Question 1 on your answer form "a".

2. The standard error of the estimated slope in the summary of a simple least squares regression equation
   Estimates the sample-to-sample standard deviation of the estimated slope.

3. The RMSE of a fitted simple regression model
   Estimates the standard deviation of ε in the Simple Regression Model.

4. A key limitation of the R^2 statistic in regression analysis is that it
   Increases when any explanatory variable is added to the regression.

5. The presence of heteroscedasticity in an estimated regression model
   Indicates that 95% prediction intervals from the model are unreliable.

6. The most useful plot for identifying the presence and size of autocorrelation in an estimated regression model is
   A scatterplot of the residuals versus the lagged residuals.

7. If the residuals from a fitted least squares regression are not normally distributed, then
   The 95% prediction intervals may be too long or too short.

Questions 8-14. Cash for Clunkers

8. The equation of the estimated model implies that
   Customers who drive more than 50,000 miles purchase cars averaging nearly 44.2 MPG. (As the number of miles driven increases, the estimate from the equation gets closer and closer to the intercept, 44.187 MPG.)

9. The shown Root Mean Square Error (4.85), assuming the SRM, indicates that
   95% of prediction errors when using this equation are between ± 9.7 *MPG*.
   (Plus or minus two SDs holds 95% of the observations in a normal distribution.)

10. Anne and Dave purchase cars in this program. Anne drives 10,000 miles farther than Dave annually. The difference in predicted miles per gallon of the car purchased by Anne compared to the car purchased by Dave (use the shown model and assume the SRM holds)
    Depends on how far Dave drives annually. (The equation implies a nonlinear relationship between miles driven and MPG of the chosen car. Hence, the difference between predicted values depends on the miles driven by Dave.)

11. To obtain a more accurate estimate of the slope in the fitted equation, 100 more observations are to be added from the same population as these. The greatest improvement in accuracy will be obtained by adding observations that
    Add variation to the explanatory variable. (The standard error of the slope is inversely proportional to the SD of the explanatory variable.)

The revised equation *Est MPG* = -50.595 + 9.04 $\log_e$(*Annual Miles Driven*) has $R^2 = 0.402$.

12. The standard deviation of the residuals of the revised equation
    6.05 MPG (The variance of the response from the first equation is $RMSE/(1-R^2) = 4.85^2/(1-0.616) \approx 61.257$. Hence, the RMSE of the residuals in the second equation is sqrt((1-0.402)(61.257)) ≈ 6.05

13. The best interpretation of the slope in the revised equation is
    Average *MPG* increases about 0.09 per 1% increase in annual miles driven. (Only the explanatory variable is on a natural log scale. See the display example in the casebook.)

14. The revised equation
    Has a lower R^2 than the original equation, indicating that the *original* equation fits better. (We can compare the r-squared statistics in this case because the two equation have a common response.)

Questions 15-26. A day trader regressed quarterly percentage change in the stock market on the percentage change in personal consumption expenditures for recreational services.

15. The correlation between percentage changes in the value weighted market index and percentage changes in the lag of percentage changes in purchases of recreational services is about
    0.49 (The square root of R^2 is 0.487.)

16. A theoretical model developed by the trader implies that expected percentage changes in the market in the current quarter are proportional to percentage changes in purchases of recreational services in the prior quarter. Assuming the SRM holds, the shown regression model
    Is consistent with this aspect of the model developed by the trader. (The 95% confidence interval for the slope 1.38 ± 2(0.34) includes 1 and the CI for the intercept includes 0.)

17. Consider quarters in which purchases of recreational services do not change. The fitted model
    Implies that the market falls about 2.3% during following quarters. (The predicted value is the intercept when the explanatory variable equals 0.)

18. Purchases of recreational services fell 0.9 percent during the *first* quarter of 2009. The shown model, assuming the SRM holds, predicts the percentage change in the market during the *second* quarter of 2009
    Will be between -19.9% to 12.8%, with 95% probability. (The 95% predicted value is about -2.31+1.384(-0.9) ≈ -3.56. Hence, the approximately 95% prediction interval is -3.56 ± 2(8.17), or about -19.9 to 12.8.)

19. Assuming the SRM, the *p*-value for the intercept estimates that if H_0: $\beta_0 = 0$ holds, then the
    Estimated intercepts of 12.34 percent of samples lie 2.31 or farther from 0. (The *p*-value is the probability of a deviation from H_0 comparable or larger than that found in the observed sample.)

20. The plot at the right shows residuals from the fitted model. From this plot, we should conclude that

    The residuals are consistent with the assumption of normally distributed errors. (The plot is a normal quantile plot with all data within the 95% limits permitted under normality.)

21. Before relying on inferences from the fitted simple regression model, it is *most* crucial that the day trader

    Plot the residuals over time to check for the presence of autocorrelation. (These data are a time series.)

The day trader added *Quarter* to the model.

22. The estimated coefficient of *Quarter[Q1]* in the expanded model indicates that when used to describe or predict market movements in the first quarter,

    The estimated intercept to use in this calculation is -0.634. (The coefficient of the categorical variable shifts the intercept by 1.59 to -2.224 + 1.590.)

23. The output shows that the addition of *Quarter* to the simple regression does not produce a statistically significant improvement in the fit of the model because

    The effect test for *Quarter* is not statistically significant. (It is inappropriate to judge the statistical significance in this context using t-statistics.)

24. The best explanation for the fact that the slope of *Lag*(*Recreation Services % Change*) in this regression is nearly identical to its slope in the prior simple regression (1.384) is that

    The average of *Recreation Services % Change* is similar in the 4 quarters. (This is equivalent to the absence of collinearity between the two variables.)

25. The day trader who developed this model suspects that percentage changes in recreational spending have a statistically significantly larger impact on future market returns during Quarter 3 and a smaller impact on future market returns during Quarter 1. In order to investigate this possibility, the day trader should

    Add the interaction of *Quarter* and *Lag(Recreational Services % Change)* to the model. (An interaction measures the change in the slope across different subsets of the data.)

26. The Durbin-Watson statistic of the expanded model is $D = 2.01$. This indicates that

    The underlying model errors are not autocorrelated. (DW is about 2 when the autocorrelation is zero.)

Questions 27-36. Pharmaceutical advertising.

27. From the scatterplot matrix on the previous page, we can see that

    The variable with the largest SD is *Samples*. (This variable has by far the largest range show in the scales. The other statements are incorrect.)

28. The best evidence to show that this multiple regression explains statistically significant variation in the number of new prescriptions written is to observe that

    The *p*-value of the *F*-statistic is less than 0.05.

29. Anne and Dave are physicians who are both 40 years old with 12 years of experience and practice in the same community. Both have individual practices (practice size = 1), were detailed 5 times for Nosorr and 4 times for Paenex, and received 10 samples of Nosorr. Anne is a rheumatologist and Dave is a pediatrician. The estimated model

    Predicts Anne will write about 44.8 more prescriptions for Nosorr than Dave. (The estimates for speciality predict an expected difference between rheumatologists and pediatricians equal to 23.68 - (-21.13) = 44.81.)

30. According to the estimated model, an increase in DTC from 5 to 10 pages of advertising would, assuming the other types of promotion are held constant, increase the expected number of new Nosorr prescriptions written by a physician by about (assume the MRM)

    28 to 38 more new prescriptions, with 95% confidence. (The 95% confidence interval for a one-unit change in DTC is $6.6 \pm 2(0.49) = 5.62$ to 7.58. The answer is 5 times this interval.)

31. If a physician receives one detail for Paenex and one for Nosorr, then the combined effect of these details on the average number of new prescriptions for Nosorr (assume the estimated coefficients in the regression are uncorrelated and that the MRM holds)

    Changes the average by -3.6 to 1.5 new prescriptions, with 95% confidence. (The estimated difference between Nosorr and Paenex details is about 1.93 - 2.98 = -1.05. The variance of the difference, assuming independence, is $0.897^2 + 0.876^2 \approx 1.572$.
    The standard error of the difference is hence sqrt(1.572) ≈ 1.25, implying that the 95% confidence interval for the difference is about -1.05 ± 2.5.)

32. If we add the two variables *Mentions Nosorr* and *Minutes Nosorr* to the multiple regression, then we can anticipate from the shown output that

    The standard error for *Details Nosorr* will increase. (The variables *Mentions* and *Minutes* are highly correlated with *Details* in the scatterplot matrix; collinearity typically increases the standard error of the correlated explanatory variables.)

33. *Years of Experience* has a positive marginal correlation with the number of new prescriptions written for Nosorr, but has a negative coefficient in the shown multiple regression. The best explanation for the change in sign of the direction of the association is

    Years of Experience is correlated with other explanatory variables. (This type of change in the sign is typical when variables are collinear.)

34. The leverage plot for *DTC* shown to the right of this question indicates that

    Some communities receive noticeably higher direct to consumer advertising. (DTC is not specific to specialties, but rather geographically determined. The clusters to the right of the leverage plot are evidently communities with higher than typical DTC spending.)

35. The scatterplot of residuals on predicted values shown to the right of this question indicates that

    The equation under-predicts the use of Nosorr by physicians who write very few prescriptions. (The zero line is far below the data at the left side of the plot.)

36. Having seen the scatterplot of residuals on predicted values (prior question), the most useful next action to take is to

    Transform the response to capture the curvature and changing variation. (This plot suggests that a log transformation would improve the model, or perhaps the analyst should consider investigating the number of NRx per detail visit.)

Questions 37-43. A regression model with the response *Samples Nosorr*

37. The fitted equation predicts that a general practitioner who is visited 10 times by sales representatives (*Details Nosorr* = 10) receives about

    163 samples. (The prediction is about 0.991 -1.663 + (15.004 +1.386)*10 ≈ 163).

38. The fitted equation indicates that general practitioners receive on average about

    16.4 samples per detail visit. (The slope is 15.004 + 1.386 ≈ 16.4.)

39. At the average level of detailing (4.25 details overall) general practitioners receive on average about

    8.0 more samples than pediatricians. (Use the comparisons table.)

40. When detailed the average number of visits (4.25), general practitioners receive on average (assuming that the MRM holds)

    Statistically significantly more samples than pediatricians. (The confidence interval given in the comparison table .0073 to 16.03 does not include 0.)

41. It has been claimed that sales representatives on average give more samples per detail visit to some types of physicians than to others. In order to test whether such differences are statistically significant requires (assume the MRM holds)

    Checking the effect test output for the interaction of specialty and detailing. (The slope for detailing varies by specialty; the effect test indicates if these differences are statistically significant.)

42. The most important diagnostic plot that should be considered next before accepting the use of the MRM with these data is

    Comparison (side-by-side) boxplots of the residuals grouped by specialty. (The regression gives different means for the groups, and the variances may differ as well.)

43. Changes in the accounting system that records the number of samples given out now express the number of samples in terms of cartons given out rather than the count of individual sample packages. Each carton of samples contains 6 individual sample packages. If the response variable in this analysis is changed from the number of sample packages to the number of sample cartons, then

    All of the *t*-statistics remain as shown. (Changing the scale of the response does not change R^2 nor the significance of the estimates (*t*-statistics) but it does change RMSE. Slope estimates and standard errors become 6 times smaller.)