Solutions for 2008 Statistics Waiver Exam

(1) A, B, C, or D as instructed.

(2) E The probability both CIs contain β_1 is less than 0.95.
The results are independent of one another, so the chance that both contain β_1 is 0.95^2.

(3) D Zero lies inside the 95% confidence interval for β_1.
The 95% CI contains those values of the parameter β_1 that are not statistically significantly different from the estimate.

(4) D The 95% confidence interval for the slope will be wider.
Setting aside a leveraged outlier reduces the variation in the explanatory variable, typically resulting in a larger standard error for the slope and consequently a wider confidence interval.

(5) A The second model allows diminishing marginal returns.
This situation is precisely that illustrated in the "Display" example covered in the notes.

(6) E The *p*-value of the associated partial *F* statistic is less than 0.05.
The proper test for the benefit of adding a categorical variable is the partial F test.

(7) A The model explains about 38% of the variation in parts assembled.
This is the definition of R^2.

(8) D The estimated slope would be closer to zero
The slope represents 3.2 parts per hour. Expressed in minutes, that would convert to a slope of only 3.2/60 = .05333 parts per minute. R^2 and RMSE would not change since the response is the same; ditto for the intercept.

(9) E 163 parts
35.58 + 3.19 * 40 = 163.18

(10) E All of the above.
These conditions all occur in a model with statistically significant slope.

(11) B The estimated intercept is consistent with management's belief.
The *t*-statistic indicates that the estimate is within sampling variation of zero.

(12) C More than 15 minutes, but not significantly more
The confidence interval for the slope, which is the expected number made per hour, includes 4.

(13) A Increase between 10.6 to 21.3 parts
The endpoints of the 95% confidence interval for the expected increase of 5 more hours of labor are

upper: 5*(3.1874899 – 2 * 0.538605) = 10.55
lower: 5*(3.1874899 + 2 * 0.538605) = 21.32

(14) E The new employee is performing comparably to his colleagues.
The RMSE of the fitted model is 17.7; the deviation of 10 is well within the range of typical sampling variation seen in these data.

(15) C The standard error of the slope and intercept would be larger.
The RMSE and R^2 would remain similar to those seen; the CI for the slope would increase by a factor of sqrt(2) (not 2).

(16) C The number of parts assembled becomes more variable with hours worked.
The variation gradually increases with the length of time worked. To judge normality would require a normal quantile plot.

(17) D The fitted data are not independent, violating the SRM.
Paired data would imply dependence between the two observations for each employee.

(18) D The RMSE in the resulting fitted model will be about 13.
Averaging reduces the SD of data by the square root of the number of averaged terms (assuming independence). Hence, we expect the new RMSE to be close to 17.7/sqrt(2) = 12.52

(19) D *Years of Experience* is positively associated with *Salary*.
The correlation is evident in the summary table. Leverage plots do not show the range of the data, but rather the range after adjusting for other explanatory variables.

(20) A Explains statistically significant variation in the salary of reporters.
The overall F-statistic can be computed from the shown R^2. It indicates that the regression explains statistically significant variation in the response (p-value less than 0.05, F = (0.7541/(1-0.7541))*(84-3)/2 = 124.2).

(21) B The intercept represents an extrapolation far from observed data.
The intercept is very negative (yet statistically significant) due to the degree of the extrapolation from the observed data.

(22) D No, because the size of the t-statistic of *Years of Experience* in this regression.
The partial slope of the variable is not statistically significant.

(23) C Those with 250 story lines earn about $12,000 more.
This is 10 times the partial slope.

(24) A Remain the same.
There's no evidence of an increase in salary by marking time unless that time is used to produce story lines.

(25) B Collinearity reduces the precision of the slope estimates.
The leverage plots show the "narrowing" of the range of the explanatory variable seen in the presence of collinear predictors.

(26) C The slope would be larger by a factor of about 2.3.
The natural log is 2.3 times the base 10 log, (ln x = 2.3 $\log_{10}$ x) so the effect of the change in logs is to multiply the slope by 2.3.

(27) D About 13 million were predicted to watch during the first week.
The intercept in a model using logs is the predicted value when the predictor is 1.

(28) C True only for the first 9 weeks.
The model is not linear because of the log transformation. The slope is the derivative of the fit, or 18.2/week. By week 10, the slope is less than 2 (million).

(29) A 1/6
The predicted value of the model is 76.4 (million) so that one RMSE above the fit is 76.4 + 7.7 = 84.1(million). The probability of a normal more than one SD or larger that its mean value is about 1/6.

(30) A The residuals are approximately normally distributed.
The figure shows a quite "normal" quantile plot. The points appear to "track" since they are the ordered residuals, sorted by size, not by time order.

(31) A Yes, the coefficient of "Season[Second]*Week" is negative and significant.
The interaction implies a change in the slope and is significant (p-value is 0.0023). During the first season, the slope is 1.9+0.9, whereas in the second the slope is 1.9–0.9.

(32) A The fitted model estimates a baseline audience of about 22.8 million in week zero.
The intercept of the implied regression model for the first season is 34.309012-11.48625 = 22.822 million.

(33) D Increased by about 1 million on average per week
The slope in the second period is the baseline slope (1.9) plus the effect of the interaction (-0.9). The interaction alone is not the slope, but rather the shift from the baseline model.

(34) B 80 million.
The fit from the equation of the model (during season 2) is
34.31 + 1.93*33 + 11.486 - 33*0.883 = 80.347

(35) E That this view of the residuals suggests no problems with the model.
None of the outliers are so extreme as to cause a problem with this amount of data. There is no *systematic* trend toward higher variance at the larger values.

(36) D Remove the confusing interaction term from the model.
An interaction term might be confusing, but this one is quite significant.

(37) C Yes, the overall F statistic is statistically significant.
The overall F test tests for the significance of the full model. Other shown properties may indicate a good model (large R^2), but not statistical significance.

(38) E The shown results do not provide an answer to this question.
The coefficient of *Sex* compares a male applicant to a female applicant who uses the same language, takes the same length of time, and uses the same number of lines.

(39) D The claim is not true for programs written in "Java."
The slope for *Lines of Code* is positive, suggesting that the longer programs take longer to generate a page. One must take the interaction into account. The interaction between *Lines of Code* and language implies, however, that the slope for Java programs is negative. Thus, for Java, *longer* programs run faster.

(40) A "C"
Figure out the fit for each, ignoring terms in common (age, sex, coding time). For example, the relevant part of the fit for the C program is (4.47-5.84)+(.023-.0116)*200 = 0.91. The predicted times for programs written in Java and Script are larger.

(41) A The categorical indicating sex of the programmer.
Because of possible collinearity, we ought not remove two predictors at once.

(42) D Requires that we assume that the population slope is zero
You have to assume the null hypothesis of no effect (slope is zero in the population) to find the p-value. The p-value is *not* a probability for either hypothesis.

(43) A The R^2 of the resulting model would be stat significantly smaller than 0.60.
The t-statistic for this slope is much larger than 2 (with p-value less than 0.05).

(44) E Inspect side-by-side boxplots of residuals grouped by *Sex*.
These boxplots show the scale of the residual variation in each group. These are assumed to be equal in the multiple regression model.

(45) C Is accurately predicted by the fitted model.
The indicated observation has residual value near zero.