Solutions for 1999 Statistics Waiver Exam

(1) A, as instructed

(2) D

The slope of the fitted model gives the change in the response per unit change in the predictor. In this case, we have 91 ($sales/hour)/(calls/hour) = 91 $sales/call.

(3) C

From the fitted model, the predicted level of sales at 100 calls is 310+9100 = 9400. The shown residual plot suggests an RMSE of about 800, as used here.

(4) D

This outlier is located near the center of the predictor and not leveraged. Though quite negative, it is compatible with the residual variation in the quantile plot.

(5) C

The outlier is not leveraged and removing it will not affect the slope. Removing it will, however, produce a small drop in the RMSE, or residual standard deviation.

(6) A

From the expression for the F ratio as $F= (n-2)R^2/(1-R^2)$, it only takes about 5 observations to make the fit significant (F=4).

(7) E

All of the explanations are appropriate.

(8) D

This is time series data, so it would be useful to add the DW test as a check for autocorrelation in the underlying error process.

(9) B

The fit explains significant variation since the t-statistic for the slope is significant.

(10) B

The constant term is the fitted value when the predictor, here log file size, is zero. For log file size to be zero, the file size must be 1 (KB).

(11) E

Since the fitted model is logarithmic, attaching the message to the larger file will result in a smaller increase in the time for transmission.

(12) C

This is most easily done from the plot.

(13) B

This is a large deviation relative to the claimed accuracy of the model, at ±2 RMSE or about ± 6.8.

(14) B

Both the intercept and slope change with increases in the number of users, thus both a term for the number of users and its interaction would be needed.

(15) A

From the positive correlation (0.54) of displacement with price. “Marginal” in the question implies that we average over the role of other factors, whereas a partial effect would adjust for them as in a regression model.

(16) A

The slope for number of cylinders indicates that each adds a cost of about $2100 to the car, so double this for two additional cylinders.

(17) A

The overall fit (from the F = 48.5) is quite significant.

(18) C

The size of this coefficient is quite hard to interpret and is due to the extreme collinearity among this collection of predictors.

(19) A

The narrow leverage plot for horsepower indicates the effect of collinearity.

(20) B

The residual plot shows increasing variation with predicted price.

(21) C

Though the intervals are rather wide, the model intervals are unnecessarily wide for the cheaper cars due to the lack of constant variance seen in the residual plot.

(22) E

This is not time series data, and the DW calculation would depend upon the arbitrary ordering of the data as stored in the underlying data file.

(23) A

The correlation of VO_2 max with the response is 0.91, so this single predictor would explain 0.91^2% of the variation in duration.

(24) B

All of the other statements about the shown scatterplot matrix are false.

(25) A

An increase of 10 in max HR produces an expected change of 10(1.29) seconds.

(26) E

The coefficient for the categorical term is half the difference in the fitted intercepts, but this effect is not significant in this fit.

(27) E

We could only be assured of a more significant overall fit.

(28) D

Since the athlete with the lower time does not stay on so long, she is doing worse. Since the RMSE of the fitted model is 51.2 sec, the difference in performance of these two women is slight given the accuracy of this model.

(29) A

The t-statistic in a multiple regression measures the significance of the addition of the predictor to the fit.

(30) C

The crosses at the right of the plot indicate larger predicted times for the men in the analyzed data. None of the other claims can be made from this plot.

(31) D

You don’t need to do the partial F when all of the added coefficients are this significant. (You can compute it here and see that it also implies an improved fit.)

(32) D

The VIF measures the impact of collinearity, and both of these estimates have been affected. Nonetheless, both are still significant.

(33) B

Changing VO_2 impacts duration both through the size of the its regression coefficient as well as the quadratic term. The quadratic term implies that the incremental effect of increasing VO_2 gets larger as the value of VO_2 increases.

(34) C

The other claims are false. VIF terms do not indicate significance, and while there's a lot of collinearity in the HR terms, both remain siginificant.

(35) D

The chi-square statistic 14.1 (or a confidence interval, 0.7 ± .4) indicates that the fit is significant.

(36) D

With 5 children, the predicted log(odds) is negative (-3.6+.7(5) = -0.1). Thus, the odds are just slightly less than 1-to-1. Alternatively, look at the plot and work out the needed probabilities directly.

(37) C

You can just compute the estimated probability from log(p/1-p) = -3.6 or read it from the plot.

(38) C

The implied probability with one child is near 0.05, and about 0.10 with two.

(39) D

The comparison of all paired means shows large differences.

(40) B

The male subjects with either college or professional exposure rated the project significantly higher than others.

(41) D (not scored)

The interaction term indicates that the effect of education differs from men to women. You can also see this difference in the group means. Males with HS education were the lowest group among the men, whereas women with HS were highest among women. (The response for D should have read "No, because of the significant interaction.")

(42) B

The multiple comparisons to the max using Hsu's method show no difference at 40 or 50 degrees, but significant differences elsewhere. Other temperatures produce smaller deposits. The question is considering the temperatures for the most extreme deposits, and thus Hsu's method is appropriate.

(43) C

Again, using Hsu's method the results at 30 are lower than at 40 degrees.

(44) B

The overall F indicates the significance of the difference among the means. The size (F=15) as well as the p-value show that the differences are large.

(45) C

There are several ways to solve this, such as the relationship of F to R^2. The easiest, however, is to use the sums of squares in the Anova table. The R^2 for this fit is thus $R^2 = 3855/6902 \approx 0.559$. A polynomial fit can't do any better than this.

(46) D

Since the variances appear quite different at the different temperatures, using the nominal procedure of fitted value ± 2 RMSE will be too wide at 30 degrees. The actual prediction errors will be shorter.