# Solutions for 2001 Statistics Waiver Exam

(1) C

At 70 degrees, the model predicts the KWH use to be –7.87+2.3(70) = 153.13, so the closest answer is 150.

(1) B

A 5 degree increase implies the expected change in KWH use to be in the range implied by the confidence interval for the slope, or 5[2.3 ± 2(.21)] = 5[1.88, 2.72] = [9.4, 13.6], or about 9.5 to 13.5 KWH.

(3) D

The slope is the implied change in use per degree change. The claimed value of 2 lies inside the confidence interval for the slope, so the claim is met.

(4) C

The prediction at this temperature is –7.87+2.3(100) = 222.13, so the model is right on target.

(5) C

The difference in KWH on two days with equal temperature requires a prediction interval for the difference of two random errors. For this one has to take the square root of twice the MSE of the fitted model, or sqrt(2 $(20.2)^2$) = 28.6. This is the SD of the difference of two model errors, so the prediction interval is twice this range.

(6) C

With 4 times as many observations, one expects the SE of the slope to drop by a factor of sqrt(4) or 2. B is not correct since the estimate is not significant. One cannot tell what will happen to this estimate when the new data is added.

(7) C or D

The current slope is 2.3 KWH/(DegreeF). Fahrenheit degrees are "smaller" than centigrade degrees, so the easy answer is to recognize that the slope would be larger with a centigrade scale. A more careful analysis reveals two correct answers (this was not intentional, but happens). When expressed on a centigrade scale, the fitted model becomes

$$\begin{aligned}\text{KWH} &= -7.87 + 2.3\ (\text{F degrees})\\ &= -7.87 + 2.3\ (\ (9/5)\text{C}+32)\\ &= (-7.87 + 2.3(32)) + ((2.3)(9/5))\ \text{C}\end{aligned}$$

so that both the slope and the intercept are larger.

(8) A

This point lies above the line and to the left of the mean value of the predictor.

(9) D

The addition of this 0/1 indicator for hot days would allow the intercept of the fitted model to change, thus permitting a shift in the fit starting at 85 degrees while keeping the slope constrained to a common value.

(10) C

That the points remain inside the indicated bounds shows that the variation in the residuals is consistent with that predicted from normality. It does not, however, prove that the data are a sample from a normal population.

(11) A

The t-statistic for the added variable is significant.

(12) A

The rate of increase of the revised model is just the derivative of the fitted model, or 2.5 + 2(.025)(Temp–81). At a temperature of 91 degrees, the revised model thus gives a slope of 2.5+ (.05)(10) = 3. This value lies outside the confidence interval for the slope implied by the linear model.

(13) E

The model explains significant variation in the response as indicated by the size of the overall F ratio (145.7) reported in the anova table.

(14) C

The slope is the profit per customer. Taking the interaction into account, the slope for those with a license is 1.16 + (.18) = 1.34 $/Customer.

(15) B

The difference in slopes measures the difference profits per customer. This difference is significant as indicated by the size of the t-statistic for the interaction term (3.09).

(16) C

Fixed costs are evidently higher for those that handle beer (judging from the figure), but these costs are eventually recovered when enough customers visit the restaurant.

(17) E

Since the model has an interaction term, we cannot assess the statistical significance of the difference in values at 0 customers.

(18) E

The few points at the right are due to the presence of only a few restaurants with high predicted values.

(19) B

One can judge the effects of outliers or leverage points and assess the need for a transformation from the shown figures.

(20) A

Repeated measures on the same restaurants are not likely to be independent of one another, implying a certain redundancy in the data. The data consequently contain less information than is implied by treating these as 80 independent values.

(21) D

The slope in a log/log model is an elasticity, so a 2% increase in the predictor in this example implies a 1.26(2) =2.52% decrease in the response.

(22) E

This data lies far from the value associated with the intercept.

(23) A

The claimed value –1 lies outside the 95% confidence interval and is thus inconsistent with this data.

(24) D

The RMSE of the log/log model is on a log scale. Without the log transformation, the values of the response are in raw $'s instead of logs and will be much larger, implying a much larger RMSE even though the $R^2$ of the models will evidently be similar. (In fact, the RMSE of such a linear model is about 7.1 and $R^2 = 0.67$.)

(25) A

At a price of $30, the model predicts the log of the number sold to be 9.07–1.26 log(30) = 4.784 or a predicted value of 119.6. At 1 RMSE higher than the fit of the model, the prediction on the log scale is 4.784+.067=4.851 and predicted value in dollars is the exponential of this value, or 127.9.

(26) E

All of these are usual considerations.

(27) C

The overall F ratio (= 6) reported in the anova summary is significant, indicating some difference among the means but not isolating the difference.

(28) D

Without randomization, confounding is always a possibility in such an observational analysis.

(29) D

A range of + or – twice the RMSE is the predictive accuracy of the fitted model.

(30) A

Since *Displacement* is a statistically significant predictor, removing it will lead to a significant decrease in the predictive accuracy of the model.

(31) C

Since this model lacks an interaction, the effect of *Displacement* is assumed to be the same for all groups.

(32) C

The partial F test computed from the models is (either from using the shown sums of squares directly or from figuring out the $R^2$ for the analysis of variance to be 18343/190048 = 0.0965) is

$$F = ((0.752–0.0965)/4)/(1–0.752)/110 = (.6555/.248)\ (110/4) = 72.7$$

which is quite significant.

(33) E

The p-value is the probability of the slope being so far from zero as observed in this regression when in fact the true population slope is zero.

(34) E

From the multiple regression (the question requires partial coefficients that adjust for the other factors), the Origin[US] slope is –10 whereas that for Europe is +8, implying a shift of 18 horsepower.

(35) E

The significance of the Effect Test (partial F) for Origin is only to indicate a difference among the 3 groups without indicating where such a difference occurs.

(36) E

These two predictors are redundant, and their slopes are difficult to interpret (note the difference in signs) without thinking of collinearity.

(37) B

We assume constant variance of the errors, so we need to check this assumption from the model residuals.

(38) D

This model assumes the absence of interaction.

(39) E

The promotion affects customers of different income levels differently, as indicated by the significant interaction in the model.

(40) B

The model indicates that the expected reaction of middle income consumers to the convenience promotion to be 51 with a standard error of Sqrt (314.77/10) = 5.61. Thus the 95% interval will be 51 ± 2(5.61).

(41) C

A profile plot compares the means of the different groups, allowing one to judge the presence of interaction.

(42) D

The Tukey-Kramer method for multiple comparisons allows the marketing firm to compare any of the averages pairwise, such as those within each income category.

(43) D

This analysis is a balanced experiment (equal counts for each combination) so the fitted regression has no collinearity. Thus, when the interaction is removed, none of the remaining slopes change and thus they capture the same variation as before. This same explained variation is compared, however, to a larger amount of error variation (the interaction did explain variation) and so the F-ratios for the two remaining main effects will be smaller.