Solutions for 1998 Statistics Waiver Exam

(1) E
R2 alone does not convey the level of significance.

(2) B
This is the slope from the fitted model.

(3) B
The residual is very small at this point, and the fit would be largely unchanged were it removed. It's the smallest observation, and thus leveraged.

(4) E
Since the RMSE is about 2 million, adding and subtracting 4 million gets the right answer. Values cannot be positive, so you might as well trim the interval at zero.

(5) C
The plots show that the data lacks constant variance (as well as the context – each observation is a sum from a varying number of stores). At the left, the variance is small and the interval too wide.

(6) D
The interval includes zero, and thus does not contradict the obvious logic.

(7) D
Compute the fitted value from the line at x=75, then add +/- 2 RMSE's.

(8) B
This question asks for the slope of the fitted line.

(9) A
The residuals are highly autocorrelated due to the seasonal nature of consumption. Histograms are not useful summaries in the presence of autocorrelation.

(10) C
Adding seasonal terms would capture most of the dependence missed by the line. A lagged variable would capture some, but the dependence is not first-order.

(11) B
Divide 50,000 by the slope of the fitted model for MS, the lower slope.

(12) D
Use the reciprocal of the smaller slope, converting from sec/char to char/sec.

(13) B
The interaction term captures the difference of the two slopes, and hence the difference in transfer rates.

(14) D
The comparison boxplots show differences in variability. Without adjusting for this violation of assumptions, it would be premature to remove coefficients.

(15) E
The t-test assumes that the two groups are comparable, but these groups are not comparable since the slopes differ for the two. (Refer back to examples of confounding and the analysis of covariance.)

(16) C
The standard error would be smaller by a factor of Sqrt(80/120) = 0.816.

(17) B
The mean for this group (or fitted value from the model – using the mean is easier) is

14.8. The SD from the fitted model (RMSE) is 14.2. Thus, we need the probability of a standard normal being more than –1.

(18) C D was also allowed, though not as good.

(19) C

Use the first row of the Hsu's table, which indicates only Tab20 and Tab10 as lower.

(20) B

These are comparisons between type within dose sizes, so we need to use Tukey-Kramer here to be safe.

(21) D

The regression would force the effect of dose to be linear, whereas Anova allows it to take an arbitrary form. Thus, the Anova fit would be better.

(22) E

It is always good to check assumptions along the way, prior to removing factors.

(23) D

The needed type of analysis would be comparable to using a paired t-test versus a two-sample test. The procedure would have to account for the dependence.

(24) B

The probability of happy customers rises with price in this example.

(25) C

Logistic regression is a linear model for the log of the odds; exponentiating gives a multiplicative fit.

(26) D

Solve for the point where the log of the odds ratio is zero, 0 = -3.2 + 0.074 Price.

(27) A

Log odds are linear, odds themselves are multiplicative.

(28) C

Since the response is on a log scale, we interpret the slope as % change in Y for a unit change in X.

(29) D

The partial-F for Race is found in the first row of the effects test.

(30) B

The coefficient for South[No-Yes] is positive; thus, for South = "Yes" the effect is negative and significant.

(31) D

With so much data, the few outliers will have negligible effect (also, they are not highly leveraged).

(32) A

If you do the partial-F "by hand" you get F=9.2, which is significant.

(33) B

Adding a correlated factor leads to a very different estimate, in this case the slope drops in absolute size from 0.039 to 0.016. The SE is about the same, and thus the p-value rises since the result is not as "far" from zero.

(34) A

The coefficient of the union by years of education interaction implies that the effect of education on income is less in unions than otherwise.

(35) C

On a log scale, twice the RMSE is the value given in "c". If you were to log it, you'd find something different. Not the best question, but it certainly can be computed from the supplied information.

(36) C

The number of observations in each group determines the width of the boxes; all three have comparable lengths vertically.

(37) D

You might want to remove *Race* as a factor, but not because of its t-stats. You'd want to use the partial F instead to make this judgment.

(38) B

This one has the bulging in the wrong direction. Refer to Tukey's bulging rule.

(39) C

A very hard one, but you can tell that this is right by noting that the added variable X2 is not significant, but an increase in R^2 from 0.8697^2 to 0.812 is.

(40) A

Huge amounts of collinearity.

(41) C

The RMSE implies too much error to attain this forecasting goal.

(42) C

The coefficient of the single lag is roughly the autocorrelation in the absence of other effects. The trend is negligible.

(43) B

The plot suggests little linear growth or decay, and the t-statistics confirm this.

(44) D

The RMSE estimates the SD of the errors around the fit. With more data, you have a better estimate of the SD of the errors, but the estimate itself would not get systematically smaller.