Solutions for 2000 Statistics Waiver Exam

(1) A, as instructed.

(2) B

16 calls per hour. Either compute the fit using the estimated regression 19.8 - 0.14(30) or more simply look at the plot.

(3) B

Drops by about 1.5. Ten times the slope is 1.4, which rounds to 1.5.

(4) C

Decreases productivity significantly. The p-value 0.0004 is quite small.

(5) E

About the same. The RMSE is 1.7, so this observations lies within 2 RMSEs of the fitted line.

(6) C

Fitting to a larger data set, as in the Utopia.jmp example, makes the standard errors smaller. However, (a) is incorrect since the effect is overstated (the interval shrinks only by about half) and (b) is wrong since the RMSE will still estimate the error SD;.

(7) A

This statement is the definition of R^2.

(8) D

Do the partial F test, F = (.35-.19)/2 / (1-.35)/(60-4) = 6.9 which is larger than our conservative threshold at 4. The t-statistic for the quadratic component is a clue that the model has improved, and the partial F calculation confirms this effect.

(9) B

The polynomial. To see this is the case, either notice the fit to the first few at the left of the plot or note the way the linear would miss the bend.

(10) E

All of these. You can tell that income is a better marginal predictor since the points are more clustered along the diagonal in that frame of the scatterplot matrix. The ellipsoidal share of the points in the plots suggests normality is reasonable.

(11) B

It is an extrapolation here (look at the scales shown in the scatterplot matrix), and one might argue the model should not be linear at the origin.

(12) C

The t-statistic in a multiple regression tests the improvement offered by adding a variable.

(13) A

The multiple regression (partial) coefficient is appropriate since the income is known to be $100,000; and using the multiple regression slope for age gives 10 times -0.047, or about -0.5.

(14) D

Collinearity among predictors frequently causes the coefficients to changes signs as other predictors are added or removed from the model.

(15) C

The multiple regression implies that the sales will be larger for younger, more affluent customers (age has a negative slope, income is positive).

(16) C

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.

(17) B

The intercept in a model using logs is the predicted value when the predictor is 1.

(18) C

The model is not linear because of the log transformation. The slope is the derivative of the fit, or 18.2/week. By week 15, the slope is less than 2 (million).

(19) A

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 is about 1/6.

(20) A

A quite "normal" quantile plot. The points appear to "track" since they are the ordered residuals, sorted by size, not by time order.

(21) A

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.

(22) D

The slope in the second period is the baseline slope (1.9) plus the effect of the interaction (-0.9). The interaction term alone is not the slope, but rather the shift from the baseline model.

(23) C

The confidence interval for the difference in the rates of growth is twice the confidence interval for the interaction, or $2(0.9 \pm 2(0.25)) = 2[.4, 1.4]$. The slope in the first period is higher by a factor of 0.9 from the baseline, and the slope for the second is lower by the same amount.

(24) E

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.

(25) D

An interaction term might be confusing, but this one is quite significant.

(26) C

The overall F test tests for the significance of the full model.

(27) C

The p-value for Sex is 0.72 and thus this term is not significant.

(28) D

The slope for *Lines of Code* is positive, suggesting that the longer programs take longer to generate a page. One must take the interaction term 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 take run faster.

(29) A

Figure out the fit for each, ignoring terms in common (age, sex, coding time). The intercept for C language programs is quite small and overcomes the interaction effect. For example, the "relevant part" of the fit for the C program is (4.47-8.02)+(.023-.012)Lines + 0.012(188) = 0.91. That for Java is larger.

(30) A

Because of the potential for collinearity, we ought not remove two predictors at once.

(31) D

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.

(32) D

The fan-shaped pattern indicates smaller variance for shorter programs and hence more accurate predictions. On average, the model's predictions are on track.

(33) D

Removing the marked point would reduce the variation in the predictor and lead to a larger SE for this slope. The RMSE will *not* decrease if we remove a term with zero residual.

(34) D

The variance of the script residuals is much smaller than the variances of the other two groups. To discard such a model would be a hasty choice.

(35) D

The p-value is about 0.0002.

(36) A or B

The precise point at which the odds are 25% occurs at salary $93,500. It looks more obvious from the plot, but solving for the odds of 1 in 3 gives

$$\log(.25/.75) = -6.71+0.06x \text{ implies } x = 93.5 \text{ (thousand dollars)}$$

(37) D

The model is nonlinear, with the most rapid increase near $110,000.

(38) C

Since this is comparison to the group with largest mean, use Hsu's method.

(39) C

Using comparisons of one breed removes variation due to differences in animals. Drawing a sample from one breed would not introduce dependence.

(40) B

Since two means are compared, this is a t-test. The interval is formed as

$$\text{estimate} \pm 2s\sqrt{1/n1 + 1/n2} = 3.55 \pm 2 s \sqrt{2/10} = 3.55 \pm 2.68.$$

(41) D

It is not reasonable to expect the second sample to be an exact copy of the first.

(42) C

The p-value for *Breed* is quite small (<.0001). The overall F only tests the whole model, not the effect of adding *Breed* to an analysis with *Manufacturer* alone.

(43) B

The interaction term (which captures a difference in preference across breeds) is not significant. Most certainly E is likely to be true, but this conclusion is not addressed by the evidence given.

(44) C

The RMSE is the square root of the MSE, so we obtain the 95% region as 2 sqrt(3.81) = 3.9. This accuracy requires knowing *both* of the factors since they are significant.

(45) E

The fit to each "cell" of the two-way design is the mean for that cell (since anova is regression with categorical terms). Since there are only 2 dogs for each combination of *Breed* and *Manufacturer*, one residual is positive and the other is just the negative.