Solutions for 2005 Statistics Waiver Exam

(1) A

A difference of 5 degrees in the temperature on two days produces an expected change in rentals equal to 5 times the shown slope, or $5 \times 20.65 = 103.95$.

(2) C

The predicted value at 80 degrees is $-1153+20.65\times80 = 499$.

(3) C

The RMSE of this model is 108 rentals. The prediction is 900-800 = 100, or 1 RMSE below the available supply of tubes. Hence, by the Empirical Rule for the normal distribution, there is about a 16% chance of running out.

(4) E

This model says little about what would happen if the temperature were at zero degrees – in which case the river would be frozen anyway.

(5) D

Each 1 degree increase leads to about 20.65 more rentals (on average). Each rental costs $10. Taking into account the confidence interval for the slope gives a range of

$$10 \times (20.65 \pm 2*1.96) = 167.30 \text{ to } 245.7$$

(6) B

Celsius degrees are larger than Fahrenheit degrees, in the sense that a 10 degree change in F° is only 5.56 C°. Thus, each degree Celsius leads to a larger effect, with the slope shifting to $1.8 \times 20.65 = 37.17$.

(7) D

R^2 is the square of the correlation between predictor and response.

(8) E

The scatter of the points above the line is larger than that below the zero line. That is, the residuals are skewed with occasional very large, positive residuals on days when the model under-predicts the demand.

(9) E

This is the normal quantile plot. Since the data cross the boundary, the residuals appear to follow a non-normal distribution.

(10) D

Without a categorical indicator, the model would not be able to accommodate the upward shift in demand that happens in recreational sales on the weekends when more would visit than during the week, regardless of the temperature.

(11) A

The data were collected over time, and none of the shown plots reveal the time dynamics of the data.

(12) D

The interaction term is not statistically significant. This lack of significance does not imply, however, that the costs of the two types are identical.

(13) C

The slope for Capacity is 0.94 ± 0.15 $/GB, a range that includes 1.

(14) D

The equation for the fit for a 400 GB internal disk is

$$45.586 - 40.836 + 0.936*400 = \$287.98$$

(15) C

The 95% confidence interval for the difference in costs is twice the interval for the Type[external] effect, or

$$(40.836 \pm 2 \times 9.619) \times 2 = 43.196 \text{ to } 120.148$$

(16) A

The RMSE of the model is \$32.65, so these prices are less than 1 SD apart.

(17) C (or with some reluctance, E)

On average, for disks of size near 400 GB (so the small interaction is negligible) external disks cost $2 \times 40.836 = \$81.672$ more to purchase than internal disks. But internal disks cost \$50 to deploy, reducing the cost difference to about \$32. Because of the use of the word "about" in E, it will be accepted as well. When shifted by \$50, the CI for the difference does include zero.

(18) E

The comparison boxplot would allow us to see the effects of grouping the residuals in order to check the assumption of equal variance.

(19) E

The RMSE of the model is about \$33, so prediction errors of more than \$100 as seen in this plot are quite noticeable and indicate that the model does not apply to laptop disks.

(20) E

All of "a"-"d" are possible consequences of this grouping by vendor.

(21) C

The slope of the multiple regression is the effect of one more year of car age on the price when comparing cars of equal mileage. The estimated value has a rather wide confidence interval $2240 \pm 2 \times 369$ \$/year.

(22) D

Removing Mileage from the model would reduce R^2 significantly because this variable is significant in the multiple regression. This also shows that the increase in RMSE would be significant. Because of the evident collinearity between *Year* and *Mileage*, the t-stat for year would be larger. Because older cars also tend to have more milage (which also reduces the resale value), the magnitude of the slope would be larger.

(23) E

We can see the correlation between these two, but do not have enough information to know the slope.

(24) B

The slope for Mileage is the drop in resale price for each mile. The confidence interval for this effect, $-0.081 \pm 2 \times 0.026$, includes -0.10.

(25) D

These two points have the largest magnitude residuals in this data. Because the $RMSE^2$ is the average squared residual, the RMSE without these would be smaller.

(26) A

The model does not include other factors, such as the presence of desirable options or visible "wear-and-tear" that reduce the value of the car.

(27) A

The R^2 of the model is 0.28, indicating that 72% of the variation remains unexplained and due to factors not represented in the model. Note that the fit of this model is significant. We cannot tell whether any one of these would significantly improve the model.

(28) D

The RMSE of the shown model is related to the SD of the response by the approximate formula (which is quite accurate for this sample size)

$$RMSE^2 \approx (1-R^2)\ Var(Y)$$

Hence, the RMSE is about Sqrt(1-0.28) = 0.85, or 85% of the SD of the response.

(29) E

The *p*-value is the probability under the null hypothesis that the population intercept is zero of observing an estimated slope 1.49 standard errors, or farther, from zero. It is not the probability of any population value.

(30) A

The slope for *Value* is 35.8466 $Ins/($5000 increase in value), with 95% confidence interval 31.67 to 40.03. The confidence interval for 1/5 of this increase is thus this interval divided by 5, or 6.33 to 8.01.

(31) A

The t-stat for this coefficient is statistically significant. The coefficient for owning is positive, indicating these households pay more.

(32) A

The cost difference between old homes and new homes is the difference in the coefficients for Age[new] versus Age[old], with the addition of the terms from the interaction:

74.52 + (80000-55276) 0.001 - (-71.01 - (80000-55276) (0.0003)) = 177.67

(33) E

The interaction effects imply that the difference among the means can be substantial. The coefficients for Age can only be interpreted once the effect of the income level is known.

(34) E

Interpretation of the coefficient for *Bedrooms* requires care because the model also includes the number of rooms and the value of the home as predictors. One cannot increase the number of bedrooms without also increasing the number of rooms as well as increasing the value of the home. One must "hold fixed" the number of rooms, as in comparing a different use of a fixed number of rooms. The estimated difference here is then about 2 times the slope, or 2 × 25.31.

(35) E

The confidence interval for the slope of Income is 0.0012 ± .0004 $Ins/$Income. However, the interaction of *Income* with *Age* implies that this slope changes with the age of the home.

(36) C

The model assumes that the effect of the value of the home does not depend upon the level of income of the household; there is no interaction between income and value.

(37) B

The loss of significance due to redundancy among predictors is a common sign of collinearity. Obviously (and seen in the correlation), *Rooms* and *Bedrooms* are correlated.

(38) C

The leverage plot for Income reveals several households with exceptionally high incomes (given other characteristics) affect the slope for this predictor.

(39) D or E

The partial F test (effect test) for the addition of 34 more predictors is

$$F = \frac{(R^2_{new} - 0.28)/34}{(1 - R^2_{new})/(2099 - 12 - 34)}$$

(The current model has 12 non-redundant predictors, and only 34 of the state indicators are needed.) Solving for the new R^2 that would make F > 4 gives $R^2 > 0.325$, so one needs to increase R^2 by more than 0.04 to make this work. However, because the addition of state leads to 34 added predictors, the F required for significance is *much less* than 4, less than 2 in fact.

(40) D

For this question, one needs to compare two predictions. The variance of the difference of two prediction errors is the sum of the variances, or

$$Var(e_1 - e_2) = Var(e_1) + Var(e_2) = 2\sigma^2$$

Thus, the effective SD to use is not the simple RMSE, but rather Sqrt(2) RMSE ≈ 460. The difference of 650 is thus only 650/460 = 1.4. A difference of this magnitude happens by chance with probability (assuming normality) of 0.16. Unusual, but not rare.

(41) D

The plot of the residuals on fitted values shows that the data become more variable with the amount of insurance purchased. An analysis of the spending per value insured, with *Haz_Ins/Value* as the response, would remove this effect. Given the large sample, the outliers have relatively small effect. The data is not a time series.

(42) C

The natural log of the response indicates that we are building a model for exponential growth, in which case the slope (times 100) indicates the percentage rate of growth.

(43) B

The key is to form the prediction interval prior to exponentiating back to the raw count scale. In log units, the 95% prediction interval for month 49 is

$$4.732 + 0.161 \times 49 \pm 2 \times 0.334 = 11.953 \text{ to } 13.289$$

Now take the Exp[] of each endpoint.

(44) B

The fitted log model produces a linear relationship. Without the log, the data would increase upward at faster and faster speed forming a convex curve. The linear fit would then be too low for the larger values.

(45) C

The scale of the shown plots makes it difficult to detect the presence of autocorrelation. Because the data are increasing in size, a plot of the residuals on the predicted values will in effect be a sequence plot.