Solutions for 2004 Statistics Waiver Exam

(1) D
The intercept of the fitted model is the expected price for Age zero.

(2) E
The 2000 model Accord is two years older than a 2002 model, so the expected difference in price is twice the estimated slope in the model.

(3) C
Plugging 10 into the equation of the fitted model gives 15.29-10*1.29=2.39.

(4) C
An 8-year-old Accord is expected to cost 15.29-8*1.29 = 4.97 thousand dollars. That's basically how much money the buyer has, so we'd expect the buyer to be able to afford about ½ of these models.

(5) C
The standard error would be smaller because of the increase in the sample size, but the other statements are not necessarily true. The R^2, for example, might be larger or smaller, depending on the true relationship between age and cost.

(6) A
The model requires a transformation (such as logs for both the predictor and response) to capture the bend that is particularly evident in the plot of the residuals.

(7) A
The confidence interval for the annual depreciation in this model is $-1.29 \pm 2 \times 0.08$, or a range of -1.13 to -1.45 thousand dollars per year. The claimed depreciation of $1,400 lies inside this interval.

(8) D
If Age is treated as categorical, then the model will fit an average to the data for each age. You cannot get a higher explained sum of squares than fitting an average to each group.

(9) A
Adding observations that add the most variation to the predictor produce the largest reduction in the standard error of the slope, and hence yield a more accurate estimate.

(10) D
The linear model is incorrectly specified. Note that you should not compare R^2's between models with different responses. Explaining variation in cost is not the same as explaining variation in the log of cost.

(11) D
The slope in a log(Y) on log(X) model is the elasticity of Y with respect to X.

(12) A
The estimated log price from the log-log model is log(price) = 3.536 – 1.025*log(4) = -2.115. Now add in ±2RMSE to capture the uncertainty (on the log scale) to get q range of $2.115 \pm 2 \times 0.165 = 1.785$ to 2.445. Now "unlog" by exponentiating each endpoint to get the 95% prediction interval exp(1.785) = 5.96 to exp(2.445) = 11.53.

(13) A
The p-value for the slope is far below the usual 0.05 threshold.

(14) B

The confidence interval for the slope is $0.60 \pm 2 \times 0.17$, or a range of 0.26 to 0.94 thousand dollars per inch of height. Multiply this range by 10.

(15) D

The fit for a man who is 70 inches tall is $79.93+70\times0.60 = 121.93$ thousand dollars. Plus or minus one RMSE gives the range $121.93 \pm 8.3 = 113.63$ to 130.23.

(16) E

The data align in columns over the integer-valued heights.

(17) E

The p-value is the probability of a slope as large or larger assuming the null hypothesis of slope zero is true.

(18) D

The given slope is 0.60 thousand $ per inch. Each inch is 2.54 cm, so the slope using cm would be smaller, $0.60/2.54 = 0.2362$. Because this change of scale only affects the predictor, not the response, the RMSE would be unchanged. R^2 is not affected by changes of scale.

(19) E

To see if the effect of *Height* varies for men and women, both a categorical variable for *Sex* and its interaction with *Height* would be needed. Adding only a categorical variable would force the slopes for both groups to be the same, and adding only an intercept would force both fits to have a common intercept.

(20) A

Try the partial F-test. If you use the smallest value, the partial F-test gives a very significant result:

$$F = \frac{(7.5 - 5)/1}{(100 - 7.5)/247} = 6.68$$

This is statistically significant because the associated t-statistic would be the square root of this value, or about 2.6.

(21) C

The square of the correlation is the R^2 in a regression, and the largest correlation with the response is 0.5. The scatterplot matrix and correlations suggest little correlation among the predictors, so collinearity is not an issue.

(22) B

The partial slope is appropriate, and we find a difference of $300*0.0216=6.48$.

(23) C

From the Anova summary of the shown model, its $R^2 = 2627.4/5949.46 = 0.4416$.

(24) A

The overall F-ratio from the Anova table is statistically significant.

(25) B

From the Anova summary, the RMSE of this model is $\sqrt{34.605} = 5.883$, or about 6%. To have 95% confidence, we have to use ±2RMSE. From the scatterplot matrix, the values of all of these predictors lie well within experience.

(26) E

This is the normal quantile plot, and the residuals lie within the confidence bands.

(27) E

No particular leveraged outliers are apparent, and the points at the edges increase the slope. There is no sign of the narrowing associated with collinearity.

(28) D

Only adding other reasonable predictors will generate an improvement in accuracy. Adding more data will reduce the error in slope estimates, but does offer the chance for substantial improvement in the accuracy.

(29) B

Plugging into the fitted model gives a prediction -39.415 - 0.092 + 0.02*2004 = 0.573. You need only use the term for Q1, ignoring the others associated with quarter.

(30) B

The difference in rates between two quarters in the same year is the difference between the corresponding Quarter terms. The coefficient for Q3 is 0.122 (SE = 0.015) and for Q4 is -0.08 (SE = 0.015). The associated confidence intervals do not overlap and are quite distinct.

(31) A

The RMSE of this model is 0.038, so a 5% shift is less than 2 RMSE's away from the predicted value of the model.

(32) E

This model does not address the conjecture, which would require an interaction between year and quarter.

(33) B

The confidence interval for the slope is $0.02 \pm 2 \times 0.006$, and so includes 0.03.

(34) D

The residual plot shows points with large predicted occupancy (Q3) have large variance. We should not pick apart the several terms added by the categorical factor, and we can tell by the significance of 3 of the 4 terms that the partial F for this effect is significant.

(35) B

The effect for union relations is the difference between the two coefficients for this factor, or 2×1.23.

(36) A

The RMSE would increase significantly because a significant predictor would have been removed, leaving substantial extra variation in the residuals.

(37) B

The categorical variable for Shift[Yes] is positive (indicating more days absent), but this effect is not significant.

(38) B

An interaction measures whether the effect of one predictor (Shift) depends upon the levels of another (Union yes or no).

(39) D

Comparison boxplots are particularly useful in models with categorical predictors for checking whether the variance of the residuals is comparable in the different groups.

(40) B

The outlier is pulling the slope for *Wage* closer to zero. Removing this point would also reduce the RMSE of the model since it has a large residual. It's hard to say what will happen to the SE of the slope because this point contributes variance to the predictor, but also inflates the RMSE of the model.

(41) E

The slope for *Wage* for a company with no shift work and poor union relations is

$$-0.14 + 0.11 + 0.09 = 0.06$$

The added terms come from the interactions between *Wage* and the indicated levels of the categorical terms.

(42) B

The fit is significantly better as both added features are statistically significant.

(43) A

Collinearity is correlation among the predictors, regardless how constructed.

(44) C

The interaction shows that the slope for wages increases in companies with good relations. Adding a positive value to the negative overall slope for *Wage* slows the reductions obtained by raising wages.

(45) A

Plugging into the equation, the predicted average level of absenteeism is

$$11.97 - 0.14*50 - 0.12*35 + 0.06*50 - 1.27 - 0.6 + (50-48)*0.11 + (50-48)*0.09$$

which is 2.3. Adding ± 2 RMSE gives a 95% prediction interval

$$2.3 \pm 2 \times 2.456 = -2.612 \text{ to } 7.212$$

The negative lower value should be set to zero because you cannot have negative days absent (the model is flawed and needs a transformation). Multiplying this range by 100 gives the indicated range.