Solutions for 2007 Statistics Waiver Exam

(1) A

Solve for the unknown number of customers x in the equation of the regression model,

$$205{,}000 = 10876 + 484 \ (x = \text{number customers})$$

which implies

$$x = (205000\text{-}10876)/484 \approx 400$$

(2) A

At $x = 500$ customers

$$\text{Estimated sales} = 10876 + 484 * 500 = \$252{,}876$$

The RMSE implies that the sales threshold \$205,000 lies $(252876\text{-}205000)/45892 \approx 1$ RMSE below the predicted value. According to a normal distribution, the chance of such values is 1/6.

(3) D

The slope is the incremental impact on the average sales for one more purchasing customer.

(4) B

The intercept is not statistically significantly different from zero. A linear equation with intercept zero indicates that y is proportional to x.

(5) D

This question asks about the slope, which is \$484 ± 2(104) per additional customer. The 95% CI includes \$500, so the estimate is less, but not by a statistically significant margin.

(6) D

The 95% confidence interval for the slope is \$484 ± 2(104.2) per additional customer, which works out to (multiply the endpoints by 1000) to \$27,600 to \$69,200.

(7) E

The correlation is the same since the new response is a multiple of the previous response (change of units). The slope and intercept would be smaller. The RMSE carries the units of the response, so it would be 0.06 times the value shown, or 0.06 * 45892 = \$2753.52.

(8) B

The 95% prediction interval, so long as not extrapolating, is about $\hat{y} \pm 2$ RMSE.

(9) E

The square root of the sample size, $\sqrt{n}$ appears in the denominator of the standard error.

(10) E

This is a typical residual plot from a model without serious problems.

(11) C

The model assumes that the errors are normal, so we check the distribution of the proxies for the errors, the residuals from the fitted model.

(12) D

The SD of the explanatory variable appears in the denominator of the SE of the slope. Adding the observations that add the most to the variation of the explanatory variable will increase the SD of the explanatory variable the most.

(13) D

The question seeks the marginal association; the correlation of assets and sales is positive and hence so is the slope in the simple regression of profits on assets. The slope in the multiple regression is negative because it fixes the type of company and level of sales.

(14) D

From the anova table, the *p*-value is quite small for the *F* statistic of the shown model. The *F*-statistic takes into account not only R^2 but also the sample size and number of explanatory variables. Because of evident collinearity, we cannot rely on individual *t*-statistics to answer this question.

(15) B

The slope would have remained the same because all variables change by the same scaling factor. This common factor cancels in the units of the slope.

(16) A

Substitute the values for Assets and Sales into the fitted equation, obtaining

$$\begin{aligned}\text{Estimated profits} &= 98.82 + 0.1154(\text{assets}) - 0.043795(\text{sales}) - 226.25\\ &= 98.82 + 0.1154 * 4000 - 0.043795*2000 - 226.25\\ &= \$246.58 \text{ million}\end{aligned}$$

(17) B

Because the model omits interactions, the slopes are constrained to be the same in both types of industries.

(18) B

Collinearity (correlation among explanatory variables) often produces changes in the direction of the effect. The partial slope need not have the same sign as the marginal correlation.

(19) D

First, double the size of the estimate in order to see the difference between the two categories. Computer companies are \$226 million below the baseline fit, and pharma companies are \$226 million above. Then, observe that the regression "controls for" the other explanatory variables in the model (sales and assets).

(20) D

The *p*-value is the probability of an estimate as far or farther from the hypothesized value (here zero) assuming the model and the null hypothesis H_0: $\beta_0 = 0$.

(21) B

An increase of \$1 billion is an increase of 1000 million, so the effect is 1000 times the confidence interval of the slope for sales:

$$\begin{aligned}1000 \times [\, 0.11544 \pm 2 \times 0.027432] &= 1000 * (0.11544 - 2 * 0.027432) \text{ to}\\ &\quad 1000 * (0.11544 + 2 * 0.027432)\\ &= \$60.576 \text{ to } \$170.304 \text{ million}\end{aligned}$$

(22) E

The distinct point in the plot denotes a company with much higher predicted sales than the other firms. Since the residual is small, it also has larger actual sales.

(23) D

This is the "cottage example" from the casebook and class notes in the context of multiple regression. Most of the variation explained by the model is devoted to separating the single large firm from the other smaller companies.

(24) C

The addition of an interaction would allow comparison of the slopes in the different industries.

(25) B

Linear growth on a log scale is exponential growth in the actual counts.

(26) B

The Durbin-Watson statistic indicates a departure from the assumption of independent in the error terms over time.

(27) D

Autocorrelation in this context refers to correlation between adjacent model errors, as estimated from the fitted residuals.

(28) A

The slope on the log scale implies a rate of growth of about 14% per week.

(29) B

In Week 53, the natural log of the number of accounts is approximately

$$\text{Estimated } \log_e \text{ accounts} = 1.6792054 + 0.1353734 * 53 = 8.8539956$$

Now remove the log by exponentiating (raising e to the shown power) to obtain

$$e^{8.854} \approx 7{,}002.3$$

(30) C

To find the length of the prediction interval, first add ± 2 RMSE to the prediction, and then exponentiate the endpoints. That gives, approximately,

$$[e^{8.854 - 2*\text{RMSE}} \text{ to } e^{8.854+2*\text{RMSE}}] = [e^{8.854 - 2*.412292} \text{ to } e^{8.854+2*.412292}]$$
$$\approx 3{,}070 \text{ to } 15{,}972$$

(31) D

Removing this period of large residual variation would produce a much smaller variation around the fitted line without increasing the SE of the slope by very much.

(32) C

Read the scale for the variable *Dollars* at the left side of the top row of the scatterplot matrix.

(33) C

The tighter the ellipse, the more concentrated along the diagonal the data become, and hence the larger the correlation between the two variables.

(34) D

The overall F statistic essentially tests the size of R^2 for the fitted model.

(35) B

The t-statistic for a slope in multiple regression tests whether the addition of the explanatory variable significantly improves the fitted model.

(36) C

The partial F statistic 52.6 (computed by JMP and labeled as the Effect Test for *PreparedBy*) is statistically significant. This test adjusts for looking at several coefficients at once, unlike peeking at individual t-statistics.

(37) B

Holding fixed the income and preparation method, the effect of $10,000 more in deductions is

$$10000 * (0.0157184 - 2 * 0.007105, 0.0157184 + 2 * 0.007105) = \$15.084 \text{ to } 299.284$$

(38) D

The predicted recovered taxes from the fitted model in category "payer" is

$$119.9538 + (0.0050454-0.001537) * \text{Income} + (0.0092495+0.0064142) * \text{Deduct} + 175.922$$
$$= (119.9538+175.922) + (0.0050454-0.001537)*100000 + (0.0092495+0.0064142)*20000$$
$$= 959.9898$$

(39) D

The partial F test for the difference in the two models is

$$(0.626981-0.616832)/(1-0.626981) * (150 - 1 - 8)/4 \approx 0.96$$

which is not large enough to indicate a statistically significant improvement in the model.

(40) E

Interactions are almost the same as the underlying explanatory variable, being a product of it with a dummy variable. The addition of such interactions adds substantial collinearity.

(41) C

The leverage plot is specifically designed to detect leveraged outliers in this context.

(42) D

The boxplots show that the interquartile ranges are similar in the three groups of residuals.

(43) E

That's a normal quantile plot of the residuals.

(44) A

Confounding means that the differences associated with the distinction among groups of a categorical variable may be due to another effect. In this case, the differences due to preparation type may actually be the result of different behaviors among the auditors.