Solutions for 2006 Statistics Waiver Exam

Questions 19 and 40 were not included in the scoring of the exam, leaving 42. For Question 26, answer "a" was judeged acceptable. For Question 35, "c" was also accepted.

(1) D
The predicted value is 119.48 + 50.52 × 5 = $372.08 thousand dollars.

(2) B
The predicted price for this home is $271,000, which is about 1 RMSE below the price suggested by this owner. According to a normal model, about 1/6 of homes list for larger than 1 RMSE above the fitted model.

(3) A
The incremental selling price goes up by about $50 per square foot. The realtor hence sees 5% of this increment.

(4) E
The intercept estimates fixed costs that do not vary with the size of the home. The average selling price is $323,000, and 119/323 ≈ 0.37. B is not correct because the slope measures the cost per additional square foot, not overall; it does not include the substantial fixed costs of a purchase.

(5) D
The 95% confidence interval is 25.26 ± 3.92 thousand dollars. The slope is the incremental price for one more square foot.

(6) C
The response would be 5% of its prior value, converting the scale of the plot to 5% of the shown range. These R^2 of the fit would be the same, but the scale of the response would be smaller, resulting in a smaller RMSE.

(7) B
The correlation between x and y is the square root of R^2.

(8) E
The estimated slope, RMSE, and R^2 would be about the same as those shown here. They might be smaller or larger. The standard error of the slope, however, would increase because of the presence of the square root of n in the standard error.

(9) E
So far, so good. The single point to the right does not indicate a problem, just shows that relatively few very large homes are sold.

(10) C
We'd have one less case (smaller n) and lose the most leveraged observation (hence have a smaller variation in x).

(11) B
Because the homes in developments are likely to be sold under similar conditions, we should question the assumption of independence. It's likely that the selling price of one home influences those of other homes in this area.

(12) C
The large homes would fill in the right side of the plot and increase the variation in x the most (given the usual assumptions). Add

(13) C

These are residuals from a regression that removes the effect of size on the selling price. Among these, those with a fireplace sell for slightly less. Evidently, homes with a fireplace have other sorts of problems – that or they just aren't popular in this area.

(14) B

We must interpret the slope as a partial, not marginal effect. (Hence, C is incorrect.)

(15) C

Fill in the regression equation, using the constant for the indicated region,

25979.9 - 4335.5 + 100 * -214.7 + 12*5349.7 = 64,370.8

(16) A

Use the difference between the region effects: 4335.5 - 8141.8. It's a decrease. We need the partial (rather than marginal) difference since the home is the same size and the price is fixed.

(17) E

The *p*-value is larger than 0.05, indicating that the estimate is within 2 standard errors of zero.

(18) D

The *t*-statistic for this slope is statistically significant, and hence R^2 would fall by a statistically significant amount were this predictor removed.

(19) B

The RMSE of the fitted model is 20,672 BTU, implying that the SE of the mean prediction error is $20{,}672/\sqrt{n}$. To have the confidence interval not include zero requires

$2 * 20{,}672/\sqrt{n} \approx 10000$, or $n \approx 17$.

(20) D

Differences in the other two explanatory variables associated with the regions is another form of collinearity.

(21) B

The confidence interval for 10 times the slope of electricity rate is 10 × [-215 ± 110]. Because this interval includes 2000, we cannot reject the claim of the colleague. We need the partial estimate because homeowners do not move location nor change the size of their house, presumably, when the mailing comes.

(22) D

This plot shows the classical effect of a lack of constant variance. Instead, the variation of the residuals increases with the amount of electricity being used.

(23) A

The model does not adjust for the harshness of the winter weather or the heat of summer. The other proposed adjustments would not be useful.

(24) C

You can determine this from the scale for the number of employees in the scatterplot matrix.

(25) C

The slope in the simple regression of log production on log robots is 0.52 with standard error 0.04. Thus, the estimate is only slightly larger (by half of a SE) than the claimed elasticity; hence, we cannot reject the claim. Just because we cannot reject H_0, however, does not prove that it's true.

(26) E

The overall *F*-test measures the statistically significance of the model, taken as a whole. it does not show the statistically significant of individual estimates.

(27) B

The partial elasticity from the multiple regression is needed rather than the marginal elasticity from the simple regression since plans do not include increasing the number of employees as well. Using the partial elasticity, a 5% increase in the number of robots implies a $0.05 \times 0.3 = 0.015$ increase, or 1.5%.

(28) C

The *t*-statistic for this slope is larger than 2 in absolute size (and the *p*-value is less than 0.05).

(29) E

The prediction interval for the output under these conditions is

$$\text{Exp}[8.6331 + .23228 \text{ Log}[300] + 0.30551 \text{ Log}[80] \pm 2 \text{ RMSE}] = [66285.5, 97904.8]$$

(30) B

You can see the effects of the collinearity in the narrowing of the data range in the leverage plot.

(31) A

The data are a time series, and we need to check for autocorrelation in the residuals.

(32) C

The average level of sales is $1491. To be within 10% for sales on a typical day, the prediction would need to be within about $150 of the actual value. The RMSE of the fitted model, however, is $209. Hence, there's less than 2/3 chance of being able to meet this goal.

(33) C

An increase of 20 gallons in sales, at the given price for the day, is associated with an average increase in sales of $20 \times (0.30 - 0.05) = \$5$.

(34) C

Plugging into the given equation gives

$$2249.285 + 779.292 + (-9.957 - 4.371)(160) + (0.301 + .05)(3000) = \$1789.1$$

(35) E

We cannot judge the size of this marginal comparison from the shown model. The average level of daily sales depends on other factors, such as the volume of gasoline.

(36) C

The interaction between *Price* and *Weekend* shows that the slope for price is $-9.9-4.37 =-14.27$ during weekends.

(37) D

Most of the prices are near $1.45, but for a few points spread out from this typical level.

(38) B

The slope for Price in the current model has units sales per 1 cent change in the price of gasoline. If gasoline were measured in dollars, the slope would be the sales per 100 cent change, or 100 times larger. The SE would also be 100 times larger, and hence the *t*-statistic and *p*-value would remain as shown.

(39) C

The interaction term is not statistically significant, so the change would not be statistically significant. We cannot tell how the slope for *Price* would change.

(40) C

The sum of the 5 prediction errors has SD given by $\sqrt{5}$ RMSE = $467.39. Now multiply by 2 to get 95% probability, assuming normality.

(41) E

We cannot tell from what is shown. Because *Car Wash* is related to the other explanatory variables, we would need to take collinearity (as well as day of the week) effects into account.

(42) A

The slope within this subset is the slope from the multiple regression for weekdays, namely -9.957+4.371.

(43) E

The variation appears very similar in the two groups, as required.

(44) C

The sequence plot of residuals shows no meandering pattern, implying that the DW statistic would be near 2 (and the autocorrelation near 0).