Statistics 621 Waiver Exam

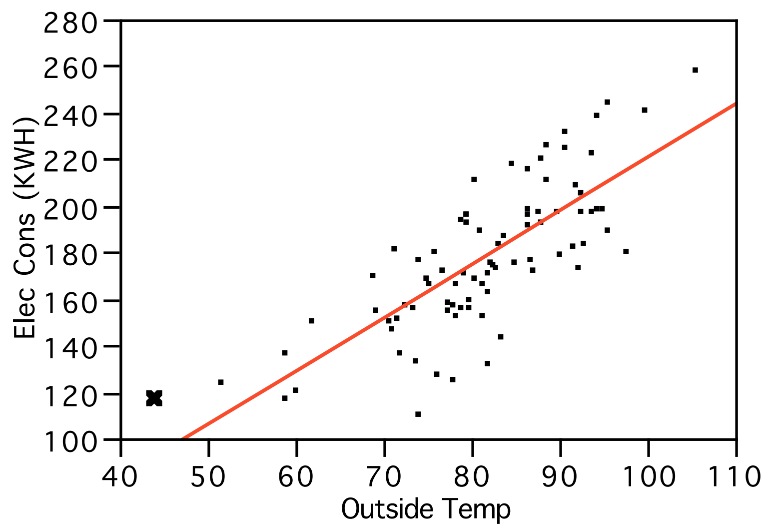
August 26, 2001

This is an **open-book** test. You have **two hours** for the exam. The computer output associated with one or more items should be considered an essential part of the questions. The multiple-choice questions are equally weighted. Please also note the following:

- **Fill in your name and student id number** on the answer form.
- **Mark the "bubbles"** under your name and student id number on the form.
- Choose **one *best* answer** by marking the item on the answer form.
- Mark the answer form using only a **#2 pencil**. Erase all changes completely.

Turn in only the answer form; you may keep the test. Solutions will be posted on the Statistics 608 web page. Mark your copy of the exam copy in order to see how well you did. The list of those passing this exam will be available in 111 Vance.

(Questions 1– 12) A small office gathered data on its energy use over a recent period during the summer months. For 85 consecutive days, the office recorded the temperature outside the building with a shaded thermometer at 12 noon. The temperature was measured in degrees Fahrenheit. (If *x* is the temperature in degrees Fahrenheit, then the corresponding temperature in centigrade is (5/9)(x-32). For the 24-hour period midnight to midnight, the office recorded its total electricity use, measured in kilowatt-hours (KWH).



Elec Cons (KWH) = -7.87 + 2.30 Outside Temp

Summary of Fit

RSquare	0.594
Root Mean Square Error	20.2
Mean of Response	178.4
Observations	85

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-7.87	17.04	-0.46	0.6454
Outside Temp	2.30	0.21	11.02	<.0001

(1) The fitted model implies that if the temperature recorded at noon is 70 degrees, then the office should anticipate its electricity use to be about
(a) 130 KWH.
(b) 140 KWH.
(c) 150 KWH.
(d) 160 KWH.
(e) 230 KWH.

(2) The temperature recorded on a Monday is 85 degrees. From the fitted model, if the temperature on the next day is predicted to be 90 degrees, then the office should expect on average to use
(a) Between 10.5 to 12.5 more kilowatt-hours on the hotter day.
(b) Between 9.5 to 13.5 more kilowatt-hours on the hotter day.
(c) Between 11 to 12 more kilowatt-hours on the hotter day.,
(d) Between 160 to 240 more kilowatt-hours on the hotter day.
(e) Between 40 to 80 more kilowatt-hours on the hotter day.

(3) The energy-conservation design of the office promised that the electricity use of this facility would rise on average by about 2 kilowatt-hours per one-degree increase in outside temperature. Does the fitted model indicate that this office meets the design?
(a) No, the negative intercept implies a poor design.
(b) No, the office uses more than this amount as the temperature increases.
(c) No, the office uses less than this amount as the temperature increases.
(d) Yes, the data are consistent with the stated design.
(e) Yes, the estimated intercept is not significantly different from zero.

(4) The recorded use of electricity was 222 KWH on a day for which the temperature at noon was 100 degrees. The model indicates that this use is
(a) Much lower than predicted, indicating better than expected performance.
(b) Lower than expected, but consistent with typical day-to-day variation.
(c) On target with the predictions from this model.
(d) Higher than expected, but consistent with typical day-to-day variation.
(e) Much higher than expected, indicating exceptionally high use for that day.

(5) On two consecutive days, the temperature at noon was 80 degrees. Based on this model, the *difference* in energy use on these two days should typically (i.e., 95% of the time) fall in the range
(a) –20 to 20 KWH.
(b) –40 to 40 KWH.
(c) –60 to 60 KWH.
(d) –80 to 80 KWH.
(e) Cannot be determined; the same temperature introduces dependence.

(6) This fitted model uses data from 85 days. If the office were to add 85 more days of data from each of the 3 prior years (and the properties of the office were the same in these years), then
(a) The confidence interval for the slope would be 1/4 as long.
(b) The t-statistic for the intercept would be twice as large.
(c) The standard error for the slope would be about half the size found in this model.
(d) The revised model would obtain a better fit, with higher R^2 and lower RMSE.

(e) The additional data could not be added since the temperatures would not be the same.

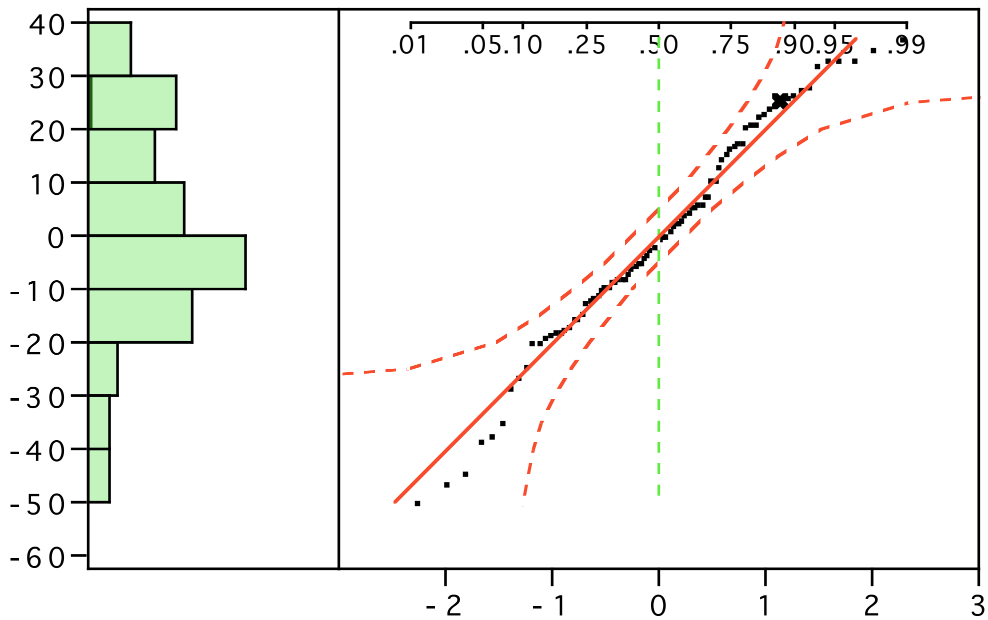
(7) If the temperature were measured in degrees centigrade rather than Fahrenheit, then (degrees C = (5/9)(degrees F – 32).
(a) The RMSE of the model would be smaller.
(b) The slope and intercept would become smaller.
(c) The slope would become larger.
(d) The intercept would become larger.
(e) The fitted model would not change.

(8) If the day with the lowest temperature at noon is excluded from the analysis (this point is marked with an X in the plot of the data), then the slope of the fitted model would
(a) Become slightly larger.
(b) Not change.
(c) Become slightly smaller.
(d) Increase to a value outside the current confidence interval.
(e) Have a smaller standard error.

(9) The office is considering adding to the model a predictor coded as 0 when the temperature is less than 85 and coded as 1 when the temperature is greater than 85. If it adds such a predictor, then the resulting model would
(a) Explain significantly more variation in KWH.
(b) Allow extrapolation to predictions of the KWH on hotter days than these days.
(c) Allow a test of whether the rate of use of KWH increases on hot days.
(d) Allow a test of whether the use of KWH increased abruptly at 85 degrees.
(e) Not be interpretable because of the introduction of collinearity.

(10) The following normal quantile plot of the residuals from the fitted linear model shows that the residuals of this model

(a) Track over time and so violate the assumption of independence.
(b) Have mean zero, indicating that the important assumptions are satisfied.
(c) Have a distribution which resembles the normal distribution.
(d) Are a sample from a normal distribution.
(e) Have a distribution which differs significantly from the normal.



Normal Quantile Plot

In an effort to improve the predictive accuracy of the model, a quadratic term was added to this model. The following tables summarize the revised model.

Summary of Fit

RSquare	0.616
Root Mean Square Error	19.802

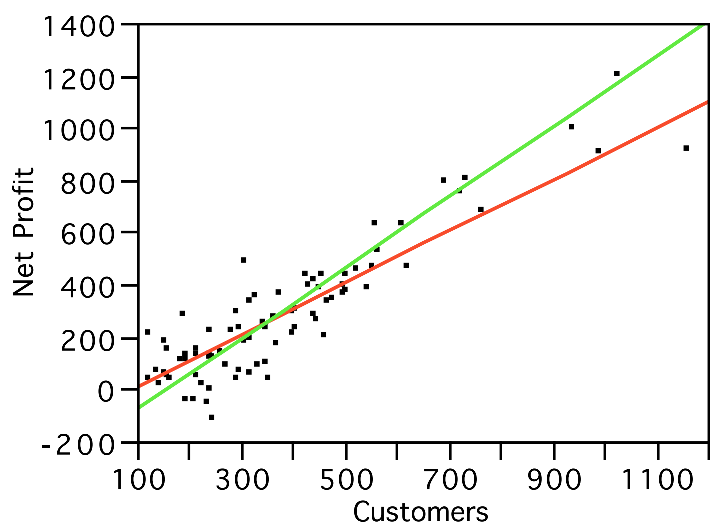
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-27.097	18.923	-1.43	0.1560
Outside Temp	2.502	0.225	11.12	<.0001
(Outside Temp-81.025)^2	0.025	0.012	2.15	0.0344

(11) Predictions from the revised model
- (a) Are significantly more accurate than those offered by the linear model.
- (b) Are less accurate than those offered by the linear model near 81 degrees.
- (c) Are slightly less accurate than those offered by the linear model.
- (d) Are not statistically different from those of the linear model.
- (e) Represent a gross extrapolation from the data and are not reliable.

(12) The revised model implies an expected increase in KWH when the temperature rises from 91 to 92 degrees that is
- (a) Inconsistent with the estimate offered by the linear model.
- (b) Consistent with the estimate offered by the linear model.
- (c) Not comparable to the estimate from the linear model since the revised model is not linear.
- (d) Inferior to that from the linear model because the quadratic adjustment is ineffective.
- (e) Significantly smaller than the estimate offered by the linear model.

(Questions 13-20) The operator of a collection of restaurants has done an analysis of past sales data to see the effects of whether or not the restaurant has a license to sell beer with meals. The data used for the following analysis gives the average daily number of customers in a recent week (*Customers*) and the net profits (*Net Profit*, in dollars) on average per day for that same week. The categorical variable *License?* indicates whether or not the restaurant possesses a license to sell beer with food; it is coded as "Yes" or "No".



Summary of Fit

RSquare	0.852
Root Mean Square Error	101.745
Mean of Response	304.149
Observations	80.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	4525367.6	1508456	145.7151
Error	76	786758.7	10352	Prob > F
C. Total	79	5312126.3		<.0001

Expanded Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-140.26	25.90	-5.41	<.0001
Customers	1.16	0.06	20.07	<.0001
License?[No]	-8.62	11.89	-0.73	0.4707
License?[Yes]	8.62	11.89	0.73	0.4707
License?[No]*(Customers-379)	-0.18	0.06	-3.09	0.0028
License?[Yes]*(Customers-379)	0.18	0.06	3.09	0.0028

(13) Does this model explain a statistically significant amount of the total variation in the net profits of the restaurants?
(a) Yes, the R-square statistic is larger than 50%.
(b) No, some of the regression slope estimates are not significant.
(c) Yes, some of the regression slope estimates are significant.
(d) Yes, the RMSE is smaller than the mean response.
(e) Yes, the overall F-ratio is significant.

(14) The fitted model suggests that the expected net profit per customer in a restaurant that has a beer license (i.e., those for which *License?* = "Yes") is about
(a) – $140
(b) $1.16
(c) $1.34
(d) $8.62
(e) $17.24

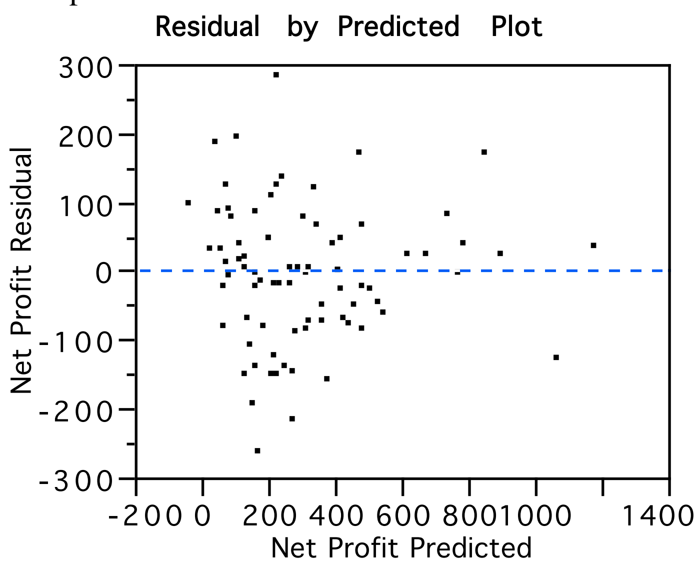
(15) Does a restaurant profit significantly more on average (in the statistical sense) per customer when the restaurant has a license to sell beer?
(a) Yes, about $0.18 *more*.
(b) Yes, about $0.36 *more*.
(c) No, the profit is essentially the same for both.
(d) No, the profit is about $0.18 *less*.
(e) This analysis cannot answer the given question; the question requires a two-sample t-test.

(16) Based on this analysis, if the cost of such a license is the same for any of these restaurants, for what type of restaurant will such a license be most profitable?
(a) Small operations with relatively few customers (say less than 200).
(b) Average operations with about 400 customers.
(c) Large operations with very many customers (say more than 600).
(d) Any of the restaurants would gain equally by purchasing such a license.
(e) None of the restaurants would gain any profits by acquiring such a license.

(17) Does the model suggest that fixed costs (those that occur without any customers, that are not proportional to the number of customers) differ between those that do have a license and those that do not?
(a) Yes, the interaction is significant.
(b) No, the interaction is significant.

(c) Yes, the difference in constant terms is significant.
(d) No, the difference in constant terms is not significant.
(e) Cannot be resolved from the output due to the interaction.

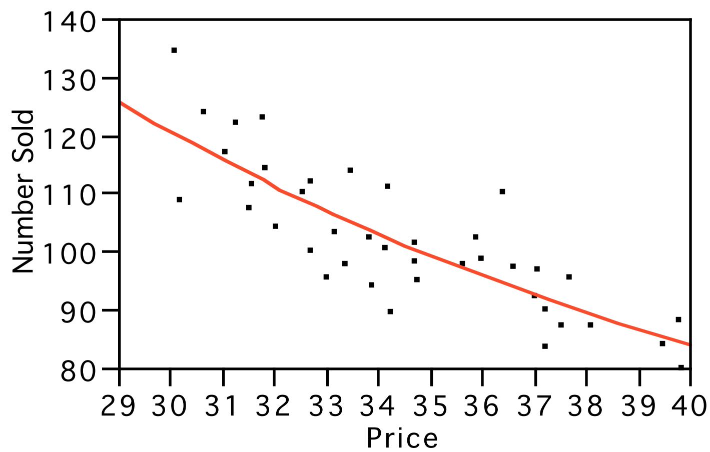
(18) The residual plot shown below for this model indicates that
(a) The variance of the errors for the two types of restaurants differs.
(b) The model requires a transformation to correct for non-linearity.
(c) The model systematically under-predicts restaurants that do not have a license.
(d) The residuals are not approximately normally distributed.
(e) There are no evident problems in the fitted model.



(19) An important further diagnostic check for this regression that is not represented in the shown output would be to
(a) Check that one had equal numbers of restaurants in both categories.
(b) Use comparison boxplots to verify the assumption of equal variance.
(c) Use leverage plots to investigate the effects of outliers.
(d) Use several transformations of the response to see if one can obtain better fits.
(e) Plot the residuals of one group on those of the other as a check for dependence.

(20) Suppose that you were to learn that these data are not from a collection of 80 restaurants in one week, but rather 20 restaurants measured over 4 weeks of activity for each. This additional information suggests that the data are likely
(a) To be dependent and claimed levels of significance are incorrect.
(b) To lack constant variance due to time trends at each restaurant.
(c) To be non-normal due to the presence of outliers.
(d) To be non-normal due to weekly trends in the profits.
(e) To be independent and the claimed levels of significance are correct.

(Questions 21-26) A business gathered data on the volume of retail sales of a product sold in its outlets. The following analysis considers how the number sold per week of this product depends upon the price at which the item was offered for sale.



Log(Number Sold) = 9.07 - 1.26 Log(Price)

Summary of Fit

RSquare	0.697
Root Mean Square Error	0.067
Observations	40.000

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	9.07	0.476	19.08	<.0001
Log(Price)	-1.26	0.121	-10.4	<.0001

(21) The fitted model implies that a 2% increase in the price of the item is expected to yield
(a) 9 more sales per day.
(b) About two and a half less sales per day.
(c) An increase of 18% in the number of sales.
(d) A decrease of about 2.5% in the number of sales.
(e) Essentially no impact on the number of items sold.

(22) Direct interpretation of the intercept in the fitted model as a predicted level of sales
(a) Implies that sales will bottom out at about 9 per day if the price is put arbitrarily high.
(b) Implies sales will bottom out at about 9 per day if the price is lowered to near zero.
(c) Implies sales average about 9 items per day.
(d) Implies sales would average above 8000 units if the price were lowered to $1.
(e) Requires a large extrapolation from the range of experience and is not justified.

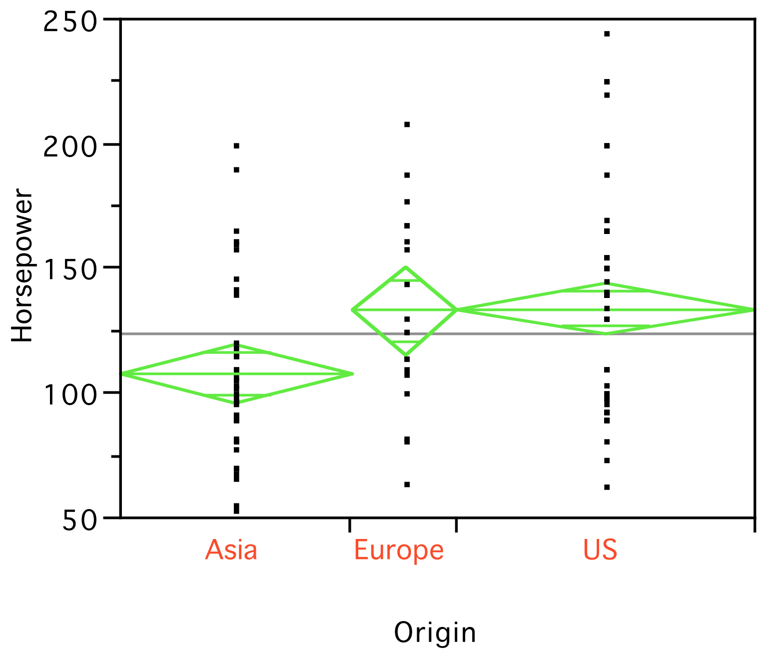
(23) It has been claimed that the elasticity of demand for this product is less than –1. With regard to this claim, the stated model implies that
(a) The claim is justified.
(b) The claim is rejected.
(c) Is plausible, but the evidence is not conclusive.
(d) More data would be required to achieve a clear answer.
(e) Cannot be answered from this analysis.

(24) Fitting a linear model that regressed *Number Sold* directly on *Price* without transformation would
(a) Have a much higher R^2.
(b) Have a much lower R^2.
(c) Have a much lower RMSE.
(d) Have a much larger RMSE.
(e) Have a smaller intercept.

(25) This model was used to predict the level of sales at a price of $30. Based on this prediction, the chance for sales to exceed 128 units is about
(a) 1/6
(b) 2/6
(c) 3/6
(d) 4/6
(e) 5/6

(26) The data for this analysis were collected over both weekdays and weekends, with the latter more busy days for the retailer. This aspect of the data suggests that one should consider
(a) If the variance of the errors is constant over these periods.
(b) If the level of sales is consistent for both weekdays and weekends.
(c) Adding a categorical indicator for weekdays and its interaction with log(price).
(d) Color-coding the observations to identify distinctions between the two periods.
(e) All of the above are appropriate.

(Questions 27-38) The following results describe several analyses of the horsepower rating of engines used in a collection of 111 different types of cars manufactured in a recent year. In the first analysis, the single explanatory factor for modeling the differences in power ratings is location of the manufacturer, either Asia, Europe, or the US. This three-level factor is denoted *Origin* in this analysis.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Origin	2	18343	9171	6	0.003
Error	108	171705	1533		
C. Total	110	190048			

Means, with Non-pooled Standard Errors

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Asia	40	107.5	6.0	95.48	119.42
Europe	19	132.9	9.0	115.10	150.69
US	52	133.9	5.3	123.39	144.50

(27) This analysis indicates that the average horsepower ratings of cars produced by manufacturers located on different continents

(a) Differ significantly between those in the US and Europe.
(b) Differ significantly between those in the US and those in Asia.
(c) Differ among the three by a significant margin.
(d) Cannot be compared because different numbers of cars are in each group.
(e) Do not different by a statistically significant amount.

(28) One should suspect the presence of confounding in this analysis because
(a) The data are clearly not independent.
(b) The groups are not balanced, having different numbers from each location.
(c) The variance of the groups is evidently different.
(d) The cars were not randomly assigned to the location groups.
(e) Several outlying observations skew the distributions for some categories.

This analysis was extended to a multiple regression. In addition to the location of the manufacturer, the multiple regression includes four additional predictors: the displacement of the engine (*Displacement*, in cubic inches), the number of cylinders in the engine (*NumCyl*), the fuel mileage in urban driving (*MPG City*, in miles per gallon), and fuel mileage in highway driving (*MPG Hwy*, also in miles per gallon). Results for the fitted model follow.

Summary of Fit

RSquare	0.752
Root Mean Square Error	20.629
Mean of Response	123.694
Observations	111

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	134504.24	22417.4	52.6784
Error	104	44257.34	425.6	**Prob > F**
C. Total	110	178761.59		<.0001

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Origin	2	2	2917.4	3.4	0.0362
Displacement	1	1	2135.6	5.0	0.0272
NumCylinder	1	1	1209.7	2.8	0.0948
MPG Hwy	1	1	1371.8	3.2	0.0755
MPG City	1	1	7367.0	17.3	<.0001

Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	121.11	25.98	4.7	<.0001
Origin[Asia]	2.63	3.10	0.8	0.3983
Origin[Europe]	7.78	4.30	1.8	0.0732
Origin[US]	-10.22	3.88	-2.6	0.0102
Displacement	0.22	0.10	2.2	0.0272
NumCylinder	6.10	3.62	1.7	0.0948
MPG Hwy	2.00	1.12	1.8	0.0755
MPG City	-5.31	1.28	-4.2	<.0001

(29) This model can
(a.) Accurately predict the horsepower for 75% of the studied cars.
(b.) Predict the power of an engine for 75% of cars to within about 20 horsepower.
(c.) Predict the power of an engine for 75% of cars to within about 40 horsepower.
(d.) Predict the power of an engine to within about 40 horsepower.
(e.) Predict the urban and highway mileage of various types of cars.

(30) If the predictor *Displacement* is removed from the fitted model, then we can see that
(a.) The R^2 of the fitted model would decrease by a statistically significant amount.
(b.) The R^2 of the fitted model would *not* decrease by a statistically significant amount.
(c.) The predictor *NumCylinder* would increase in the absence of collinearity.
(d.) The RMSE of the fitted model would increase slightly.
(e.) The differences among the slopes for *Origin* would increase.

(31) If we compare the effect of *Displacement* on *Horsepower* for cars made in Asia to the effect for cars made in the US, the fitted multiple regression implies that
(a.) Displacement has a larger effect for Asian cars.
(b.) Displacement has a larger effect for US cars.
(c.) Displacement has the same effect for both types of cars.
(d.) Displacement does not affect the power of either type of car.
(e.) Displacement significantly affects the power of US cars but not those from Asia.

(32) Does this multiple regression model explain significantly more variation in the horsepower ratings of these cars than the prior one-way analysis of variance model?
(a.) No, because too many of the added predictors are not significant.
(b.) Yes, because the slope for *MPG City* is significantly different from zero.
(c.) Yes, because a partial F-test shows a significant change in explained variation.
(d.) One cannot compare the amount of variation explained by an Anova to a regression.
(e.) One cannot answer this question from the information in the shown output

(33) The p-value for the variable *NumCylinder* indicates that
(a.) The number of cylinders has no effect on the power of an engine.
(b.) The number of cylinders has no effect on horsepower for almost 10% of these cars.
(c.) The probability that the slope for *NumCylinder* is zero is about 0.10.
(d.) The variable *NumCylinder* explains about 10% of the variation in horsepower.
(e.) None of the above.

(34) On average, six-cylinder cars made in the US with 2 liter engines that get 30 MPG highway and 20 MPG city
(a.) Have about 111 horsepower.
(b.) Have about 121 horsepower.
(c.) Have about 134 horsepower.
(d.) Have about 10 fewer horsepower than a comparable European car.
(e.) Have about 18 fewer horsepower than a comparable European car.

(35) When adjusted for differences in displacement, number of cylinders, and mileage, does this analysis indicate that cars in the US have *significantly* less power on average than those made in Europe?

(a.) Yes, the partial F-test for Origin is significant.
(b.) No, the partial F-test for Origin is not significant.
(c.) Yes, the estimated coefficient for Origin[US] is significant.
(d.) No, the estimated coefficient for Origin[Europe] is not significant.
(e.) This question requires a multiple comparisons analysis of the differences.

(36) The slopes for *MPG City* and *MPG Hwy* indicate that

(a.) Cars with powerful engines get significantly less mileage in urban driving.
(b.) Increasing the highway mileage yields a significantly more powerful engine.
(c.) The fit has been influenced by the presence of a subset of leveraged observations.
(d.) The difference in these factors is an important predictor of engine performance.
(e.) That collinearity has distorted the estimation of these slopes.

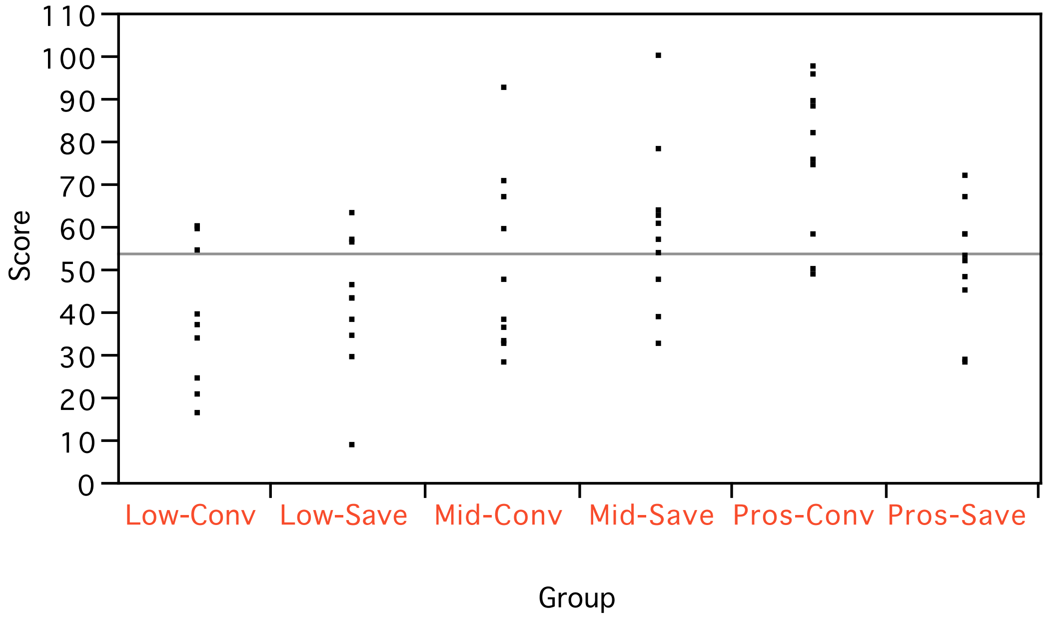
(37) Which of the following plots would be most useful in assessing the assumption of equal variation for the three groups of observations (Asian, European, and US)?

(a.) Comparison boxplots of *Horsepower* by *Origin*.
(b.) Comparison boxplots of the residuals of this model by *Origin*.
(c.) A color-coded scatterplot matrix showing *Horsepower* and these predictors.
(d.) The leverage plot for *Origin*.
(e.) A normal quantile plot of the residuals from the fitted model.

(38) A reasonable next step in the development of this regression model would be to

(a.) Check the normality of the *Horsepower* data.
(b.) Check the normality of the *Horsepower* data when grouped by *Origin*.
(c.) Remove the *Origin*[Asia] effect since it is not significant.
(d.) Consider if the effect of *Displacement* is consistent for the groups defined by *Origin*.
(e.) Check for differences in mileage among the groups defined by *Origin*.

(Questions 39–43) A marketing firm conducted an experiment to judge consumer reactions to a kitchen helper – a product designed to combine the features of several kitchen gadgets into one that would occupy a smaller amount of shelf space. Sixty consumers were grouped into three equal-sized categories, based on income (*Income*, coded as low, mid, and pros for low income, middle income and prosperous). Each group of consumers was randomly divided, with 10 from each group shown a different set of promotional material. One set of promotional material (*Promo* = conv) emphasized the space-saving convenience of the new product; the other set emphasized that such a product saved money by combining several products into one (*Promo* = save). The firm measured consumer reactions using a scored questionnaire; reactions could range from very unfavorable (*Score* = 0) to very positive (*Score* = 100).



Level	Number	Mean
Low-Conv	10	41.1
Low-Save	10	42.6
Mid-Conv	10	51.1
Mid-Save	10	60.1
Pros-Conv	10	76.6
Pros-Save	10	51.6

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	8554.229	1710.85	5.4352
Error	54	16997.687	314.77	Prob > F
C. Total	59	25551.916		0.0004

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Income	2	2	5024.9720	7.9819	0.0009
Promo	1	1	350.5061	1.1135	0.2960
Promo*Income	2	2	3178.7505	5.0493	0.0098

(39) Does the type of promotion affect the reaction of consumers?
(a.) No, the effect test for *Promo* is not significant.
(b.) Yes, the effect test for *Promo* is significant.
(c.) Yes, the overall F Ratio for the analysis is very significant.
(d.) No, the interaction term in the model is significant.
(e.) Yes, *Promo* has different effects for different income groups.

(40) From the fitted model, the average reaction of middle income consumers to the promotion that emphasizes convenience lies in the range (as a 95% interval)
(a.) 51 ± 6
(b.) 51 ± 11
(c.) 51 ± 18
(d.) 51 ± 36
(e.) Cannot be identified because the RMSE of this model is not given.

(41) To graphically display the strength of any interaction in this model, one should use a
(a.) Normal quantile plot.
(b.) Comparison boxplots shown side-by-side.
(c.) Profile plot.
(d.) Plot of the residuals on the fitted values from the model.
(e.) Leverage plot.

(42) A Tukey-Kramer analysis of the averages of the six groups in this problem would allow the marketing firm to
(a.) Identify whether the assumption of constant variance is violated.
(b.) Check for influential outliers that would otherwise distort the analysis.
(c.) Reduce the chances that the responses are dependent.
(d.) Compare the effects of the two types of promotion for each income group.
(e.) Confirm that other predictors are not confounded with the observed group differences.

(43) If the term labeled "Promo*Income" is removed from this model, then the F Ratio for *Promo* will
(a.) Not change.
(b.) Change.
(c.) Become larger.
(d.) Become smaller.
(e.) Be affected in a way that cannot be determined from the output shown here.