Statistics 608-621 Waiver Exam

August 22, 2006

On the answer sheet *before* the exam begins …

- Use a **#2 pencil**. Erase any changes completely.
- **Fill in your name and Penn student id number**.
- **Mark the "bubbles"** under the letters of your name *and* student id number on the form. Failure to do so will lead to a score of zero.

Once the exam begins …

- Choose the **one *best* answer** for each question. Picking more than one answer is scored as an error.
- You may consult **1 page of notes** during the exam. No other reference materials are permitted.
- You may use a **calculator**, but no laptops or computers are allowed. You may not use a cell phone during the exam for any purpose.

Turn in the solution page only; keep the 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.

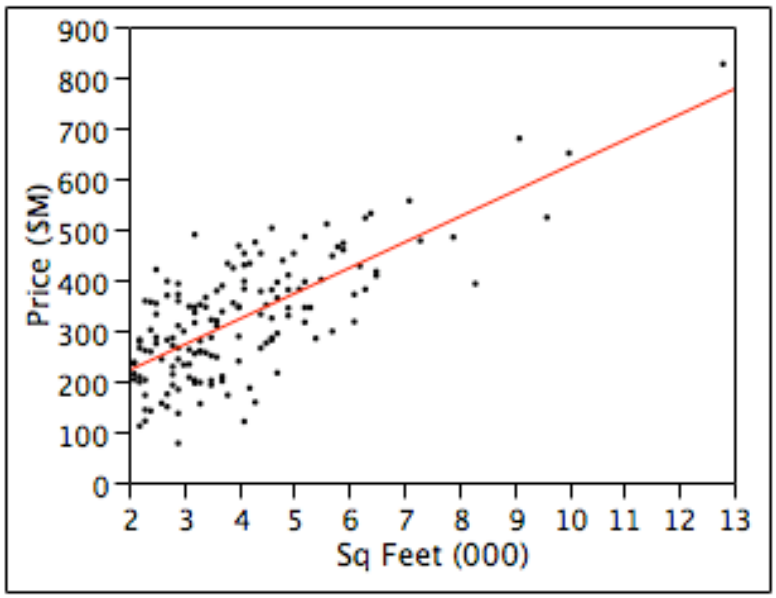
Your **score** is the number of correct answers. The multiple-choice questions are equally weighted. Some questions may be dropped and not counted as part of the overall score. There is no deduction for incorrect answers.

The solutions will be posted in WebCafé. If you wish to compare your answers to the solutions, then mark your choices on your copy of the exam. Regardless of what you write on your copy of the exam, however, only the answers marked on the grade form will be considered. You can use the "My Grades" feature to find your score as determined by your answers on the answer form.

STOP

*Do **not** turn the page until you are instructed.*

(Questions 1–13) A realtor records the size (in thousands of square feet) and selling price (in thousands of dollars) of homes that her firm has sold. In her area, the realtor is paid a 5% commission on the selling price. For example, a home that sells for $200,000 nets the realtor $10,000. The data shown in the following analysis describe the size and selling price of 150 homes.



RSquare	0.53
Root Mean Square Error	81.2
Mean of Response	323
Observations (or Sum Wgts)	150

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**
Intercept	119.48	17.12	6.98	<.0001
Sq Feet (000)	50.52	3.92	12.88	<.0001

(1) Based on the fit of this model, homes with 5,000 square feet sell on average for
- (a) $119,480
- (b) About $250,000
- (c) About $323,000
- (d) About $370,000
- (e) More than $500,000

(2) A client would like to offer his 3,000 square foot home for $350,000. According to the fit of this model, this price is (accepting the usual assumptions of the SRM)
- (a) Unusually high.
- (b) High, but exceeded by about 1/6 of homes of this size.
- (c) Typical for homes of this size.
- (d) Low, but about 1/6 of homes of this size sell for less.
- (e) Extremely low.

(3) The slope of the fitted model implies that a realtor, on average, earns about
   (a) $2.50 more per additional square foot sold.
   (b) $50 for each home sold.
   (c) $50 more per additional square foot sold.
   (d) $25 more per additional square foot sold.
   (e) The same for any sale, regardless of the square footage of the property.

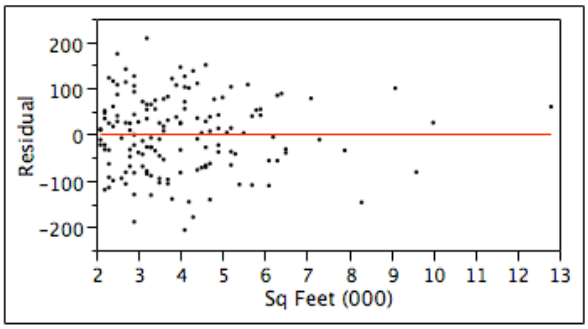
(4) This model can be interpreted to indicate that
   (a) A vacant lot sells for about $120,000.
   (b) Homes sell for about $50 per square foot.
   (c) Square footage is not related to the average selling price.
   (d) The selling price can be predicted to within less than $50 from the square footage.
   (e) About 1/3 of the average selling price is present regardless of square footage.

(5) When comparing the value of a home with 4,000 square feet to an otherwise similar home with 4,500 square feet, we'd expect the larger home to sell on average for
   (a) about the same price as the smaller home.
   (b) exactly $25,260 more than the smaller home.
   (c) approximately between $17,000 and $33,000 more, with 95% confidence.
   (d) approximately between $21,000 and $29,000 more, with 95% confidence.
   (e) approximately between $43,000 and $58,000 more, with 95% confidence.

(6) Had this regression been shown with the realtor's commission from the sale of a home on the *y*-axis rather than the selling price, then
   (a) The R^2 of the model would have been larger.
   (b) The R^2 of the model would have been smaller.
   (c) The RMSE would have been smaller.
   (d) The RMSE would have been larger.
   (e) None of the above.

(7) The R^2 of the shown model implies that
   (a) The model accurately predicts more than ½ of the home prices.
   (b) The correlation between predicted price and actual price is about 0.73.
   (c) More than ½ of the home prices are within $81,200 of the predicted price.
   (d) More than ½ of the variation in the response is unexplained by the model.
   (e) The fit of this model is not statistically significant.

(8) Had this regression been based on a fit to a sample of only 75 homes (rather than 150), we can be sure that
   (a) the RMSE of the model fit to 75 would be larger than 81.2.
   (b) the R^2 of the fitted model would be larger than 0.53.
   (c) the R^2 of the fitted model would be less than 0.53.
   (d) the estimated slope would be less than 50.52.
   (e) the standard error of the estimated slope would be larger than 3.92.



(9) This plot of the residuals from the fitted model indicates that
- (a) The selling prices of the homes are not independent.
- (b) The variance of the prices increases with the size of the home.
- (c) The variance of the prices decreases with the size of the home.
- (d) An outlier distorts the fit of the model.
- (e) The estimated model appears to meet the usual assumptions.

(10) If the largest property were excluded from the fit of the model, we can be sure that the
- (a) R^2 of the model would increase.
- (b) Estimated slope would increase.
- (c) Standard error of the slope would increase.
- (d) Estimated intercept would decrease.
- (e) RMSE would increase.

(11) For convenience, the realtor chose these 150 homes from sales in 5 housing developments. This aspect of the data suggests that the fitted model
- (a) Should not be linear.
- (b) Violates the assumption of independence.
- (c) Violates the assumption of equal variance errors.
- (d) Violates the assumption of normality.
- (e) Should be OK.

(12) The realtor plans to add prices of 10 more homes. Which choice produces the narrowest confidence interval for the slope when combined with these 150 homes (given the usual assumptions)
- (a) Pick the additional 10 homes at random from the same population as these 150.
- (b) Pick the additional 10 homes to all have 3,500 square feet (the mean size of these).
- (c) Pick the additional 10 homes from those with more than 6,000 square feet.
- (d) Choices (a)-(c) produce equal reduction in the length of the confidence interval.
- (e) We cannot identify which choice produces the narrowest interval from this information.

(13) This table summarizes the mean and SD of the *residuals* from the fitted model, grouped by whether on not the home has a working fireplace.

Has Fireplace?	Number	Mean	Std Dev
No	73	7.4694	77.9968
Yes	77	-7.0814	83.4822

This table suggests that, on average,
- (a) Homes with a fireplace sell for slightly less than those without a fireplace.
- (b) Homes with a fireplace are somewhat smaller than homes without a fireplace.
- (c) For homes of a given size, those with a fireplace sell for slightly less.
- (d) For homes of a given selling price, those with a fireplace are slightly smaller.
- (e) There's no statistically significant difference in size between the two types of homes.

(Questions 14-23). An economist developed the following model to understand how the price of electricity influences the amount used by residential customers. As the response, his model uses data for the annual electricity consumption, measured in BTUs. Explanatory variables in the model are the price per thousand kilowatt hours of electricity, the total number of rooms in the residence, and the location of the home (*Region*). *Region* is coded as 4 integers: "1", "2", "3", and "4". All 400 residences in this analysis use electricity and have central air conditioning.

Marginal Summary

Region	Number of Residences	Mean Electricity BTU	Std Dev Electricity BTU
1	81	26532.3	17332.8
2	89	31570.9	17750.3
3	180	52842.8	27738.3
4	50	37003.8	24496.3

Regression

RSquare	0.376
Root Mean Square Error	20672.03
Mean of Response	40802.05
Observations (or Sum Wgts)	400

Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	25979.9	6326.2	4.11	<.0001
Electric Rate ($/1000 KWH)	-214.7	54.8	-3.92	0.0001
Number of Rooms	5349.7	550.2	9.72	<.0001
REGION[1]	-4335.5	2775.0	-1.56	0.1190
REGION[2]	-8141.8	2023.0	-4.02	<.0001
REGION[3]	12393.3	1895.9	6.54	<.0001
REGION[4]	84.0	2381.6	0.04	0.9719

(14) The fit of the regression model implies that, on average,
- (a) Electricity is most expensive in Region 3.
- (b) At a given price and home size, residents of Region 3 use the most electricity.
- (c) Residents in Region 3 use the most electricity.
- (d) 97% of homes in Region 4 are typical.
- (e) Price has no effect on the use of electricity.

(15) For a home with 12 rooms and electricity priced at $100 per 1000 kilowatt hours, this model predicts the annual use of electricity in Region 1 to be
- (a) 12,477 BTU.
- (b) 26,532 BTU.
- (c) 64,371 BTU.
- (d) 68,706 BTU.
- (e) More than 75,000 BTU.

(16) If the home described in Question 15 were instead located in Region 2 rather than Region 1, we would expect on average its use of electricity to
(a) Decrease by 3806 BTU.
(b) Increase by 3806 BTU.
(c) Stay the same.
(d) Decrease by 8141.8 BTU.
(e) Increase by 5039 BTU.

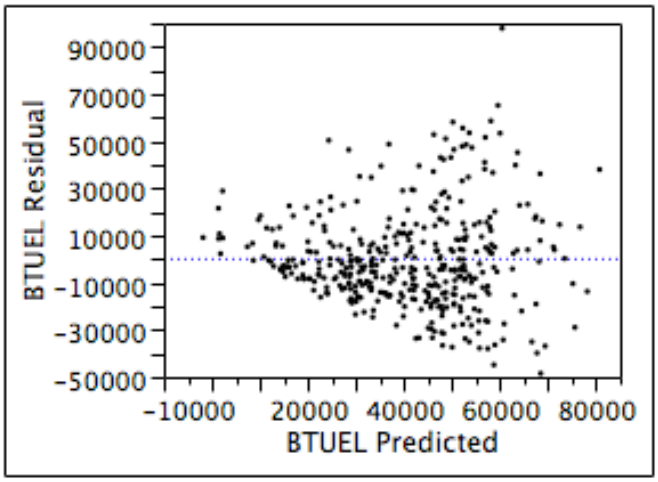
(17) The *p*-value for the coefficient of Region[1] in the shown regression output implies that
(a) 11.9% of homes in this region use the overall mean BTUs of electricity.
(b) 11.9% of samples have an estimated slope farther from zero than -4335.5.
(c) 11.9% of samples have estimated slope zero.
(d) there is an 11.9% chance that the true slope for this effect is zero.
(e) the 95% confidence interval for this estimate includes zero.

(18) If the explanatory variable *Electric Rate* were removed from the model, then
(a) The R^2 of the model would increase.
(b) The R^2 of the model would remain as shown.
(c) The R^2 of the model would decrease, but not by a statistically significant amount.
(d) The R^2 of the model would decrease by a statistically significant amount.
(e) We cannot tell because of the possible effects of collinearity.

(19) The economist believes that mailing residents information about global warming will reduce their use of electricity. He plans to send some residents this information, then compare their use of electricity to predictions from this model. If the mean use of electricity decreases by 10,000 BTU on average, then the minimal sample size needed to detect such a decrease using a 95% confidence interval is approximately
(a) 1 home.
(b) 20 homes.
(c) 100 homes.
(d) 500 homes.
(e) Such an effect is too small to be detected with predictions from this model.

(20) Mean use of electricity in Region 2 is marginally higher than that in Region 1, but the regression coefficients for these regions fall in the other order, with the estimate for Region 2 smaller than that for Region 1. This reversal is best explained by noting that
(a) The estimate for Region 1 is not statistically significantly different from zero.
(b) A leverage point has distorted the fit of the regression model, reversing the order.
(c) The model omits an important variable.
(d) Homes in different regions are of different sizes with different electricity prices.
(e) The marginal difference in mean electricity use in these regions is small.

(21) A colleague of the modeler claimed that an increase in the price of electricity by $10 per 1000 KWH would reduce the use of electricity by a typical homeowner by 2000 BTUs. According to the fit of this model, at 95% confidence,
(a) We can reject this claim; the amount used would drop by more than 2000 BTU.
(b) We cannot reject this claim.
(c) We can reject this claim; the amount used would drop by less than 2000 BTU.
(d) The model must be refit using logs to address how changes in price affect consumption.
(e) This question requires a simple regression analysis to obtain the marginal slope.



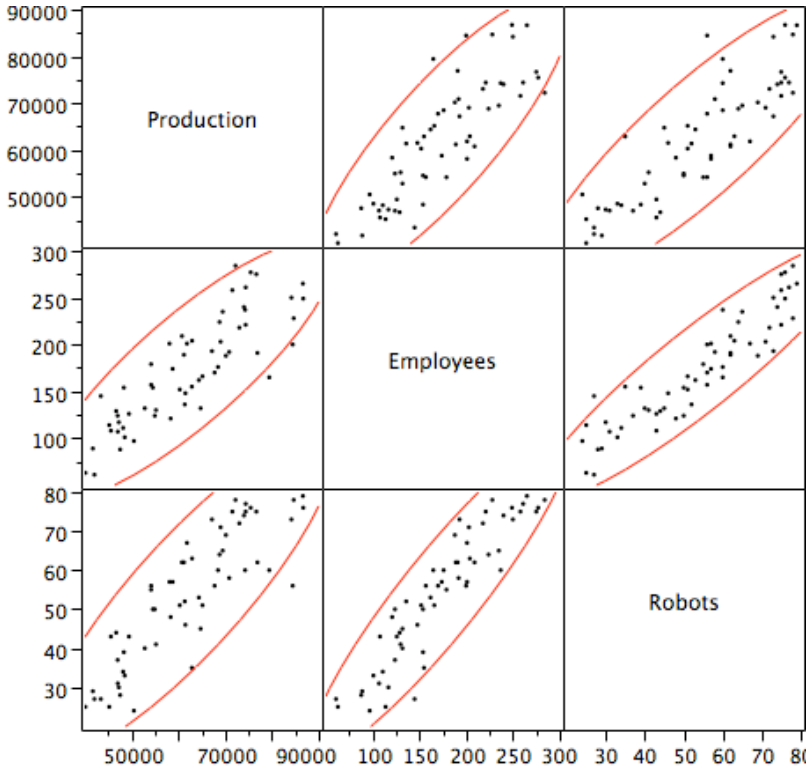
(22) The most useful interpretation of this plot is to observe that it indicates that
- (a) The data may not be independent.
- (b) Electricity use is more variable in some regions than others.
- (c) The model requires an interaction.
- (d) Electricity use becomes more variable with the quantity used.
- (e) The use of electricity in these homes is not normally distributed.

(23) An appropriate next step in this model would be to
- (a) Add an explanatory variable that measures the temperature of the locations.
- (b) Remove evident outliers seen in the previous residual plot.
- (c) Remove explanatory variables from the fit that are not statistically significant.
- (d) Represent *Region* using only 3 dummy variables.
- (e) Any of the above.

(Questions 24-31) A manufacturer that produces home appliances has steadily scaled up the size of its assembly factory. This factory uses robots to automate the production of appliances. Over the last 5 years, the company has tracked monthly of the number of robots in use (*Robots*) and the number of employees engaged in the production (*Employees*, not counting those in management). The company also tracked the shipped volume (*Production*). The following output summarizes 3 regression models: simple regressions of the log of production on the log of the number of robots and the log of the number of employees, and a multiple regression of the log of the production on both explanatory variables. All logs are natural logs.

Correlations

	Production	Employees	Robots
Production	1.0000	0.8479	0.8662
Employees	0.8479	1.0000	0.9207
Robots	0.8662	0.9207	1.0000



Simple Regression Models, y = Log(Production)

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	8.459204	0.198471	42.62	<.0001
Log(Employees)	0.5020776	0.038877	12.91	<.0001

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	8.9687773	0.151486	59.21	<.0001
Log(Robots)	0.5212021	0.038415	13.57	<.0001

Multiple Regression Model, y = Log(Production)

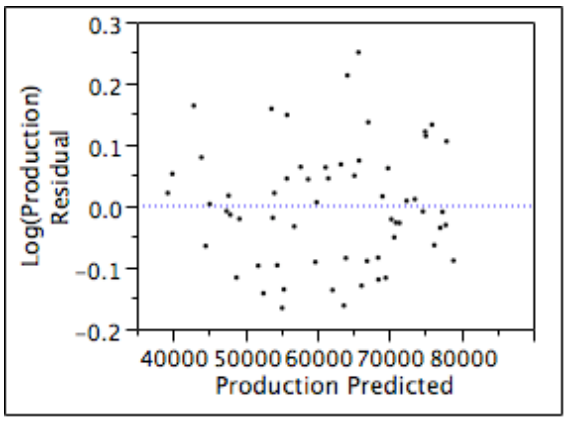
RSquare	0.788992
Root Mean Square Error	0.097506

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	2.0263521	1.01318	106.5662
Error	57	0.5419262	0.00951	**Prob > F**
C. Total	59	2.5682783		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	8.6331224	0.187511	46.04	<.0001
Log(Employees)	0.2322805	0.083603	2.78	0.0074
Log(Robots)	0.3055068	0.085729	3.56	0.0007



(24) From this output, the average number of assembly employees during these years is
(a) Fewer than 50
(b) About 100.
(c) About 200.
(d) About 300.
(e) Cannot be identified from the shown output.

(25) An economist claims that the marginal elasticity of production with respect to the use of robots of this type is equal to 0.5 in factories such as this one. The results shown here
(a) Indicate that the marginal elasticity is statistically significantly larger than 0.5.
(b) Indicate that the marginal elasticity is statistically significantly less than 0.5.
(c) Do not reject this claim.
(d) Prove that this claim holds.
(e) Do not address the claim of the economist.

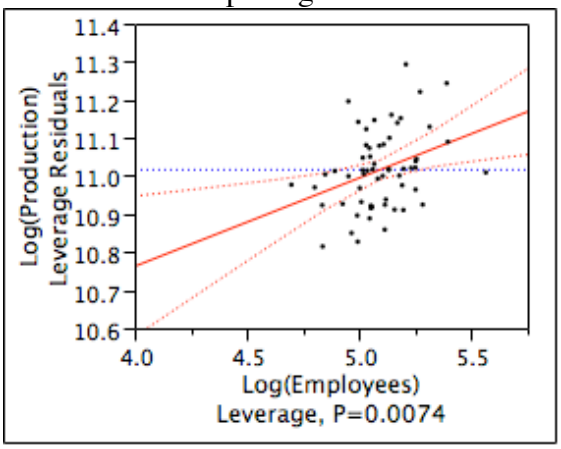
(26) The overall *F*-statistic for the multiple regression implies that
(a) At least one of the explanatory variables is statistically significant.
(b) Both explanatory variables have statistically significant t-statistics.
(c) The fitted model meets the usual assumptions of the MRM.
(d) The RMSE of the model is statistical significantly different from zero.
(e) The R^2 of the model is statistical significantly larger than zero.

(27) A manager has proposed increasing the number of robots in use from 80 to 84 while maintaining other conditions in the factory. Under the usual assumptions, these results suggest that this increase will result in, on average,
(a) A 2.5% increase in production above current levels.
(b) A 1.5% increase in production above current levels.
(c) An increase in production of 2.5 thousand units.
(d) An increase in production of 1.5 thousand units.
(e) An increase in production of 1.5 more units.

(28) The addition of Log(*Robots*) to a simple regression that contains Log(*Employees*) results in
(a) Too much collinearity to produce a model that is interpretable.
(b) A change in R^2 that is too small to be statistically significant.
(c) A change in R^2 that is statistically significant.
(d) An increase in the RMSE of the fitted model.
(e) An increase in R^2, but we cannot justify its statistical significance from the output.

(29) Currently the plant operates with 80 robots and 300 employees on the assembly line. Management has promised orders for the month of 95,000 items. The probability that production will meet this target is approximately
(a) Near one.
(b) 83%
(c) 50%
(d) 17%
(e) 2.5%

(30) The following leverage plot from the multiple regression indicates that



(a) The slope for *Log(Employees)* is not statistically significant.
(b) Collinearity has reduced the accuracy of the slope for *Log(Employees)*.
(c) Several outliers increase the size of the RMSE of the fitted model.
(d) Several outliers have pulled the slope for *Log(Employees)* toward zero.
(e) The variation of the residuals increases with the size of the factory.

(31) An important diagnostic that is omitted from this output but should be considered *next* before continuing with this analysis is
(a) A sequence plot of residuals with the Durbin-Watson statistic.
(b) Comparison boxplots of the residuals to check for equal variation.
(c) Scatterplots of the residuals vs the explanatory variable for each simple regression.
(d) A scatterplot of Log(*Employees*) versus Log(*Robots*).
(e) Histograms of each of the variables used to fit these models.

(Questions 32-44) A company that operates convenience stores collected data regarding daily sales. This analysis considers data for 285 days at one store. The response in the following analysis is the daily amount of sales (measured in dollars) at the store. This particular store also sells gasoline. (Sales of gasoline are not included as part of the response – just sales of things like drinks and snacks in the store.) Other variables known about this store each day are:

Volume (Gallons)	Amount of gasoline sold at the store.
Price (cents)	Price of the gasoline each day, in $0.01. For example, gas costing $2 per gallon would be 200 cents per gallon.
Weekend	"Yes" for Saturday and Sunday; "No" otherwise.
Car Wash	Number of cars that used the car wash at the convenience store.

Correlations

	Sales (Dollars)	Volume (Gallons)	Price(cents)	Car Wash
Sales (Dollars)	1.0000	0.8236	-0.1492	0.2085
Volume(Gallons)	0.8236	1.0000	-0.0454	0.1896
Price(cents)	-0.1492	-0.0454	1.0000	-0.0280
Car Wash	0.2085	0.1896	-0.0280	1.0000

Summary of Multiple Regression

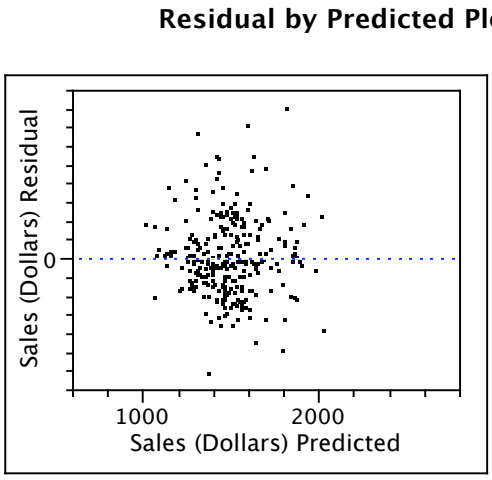
RSquare	0.444
Root Mean Square Error	208.6
Mean of Response	1490.9
Observations	285

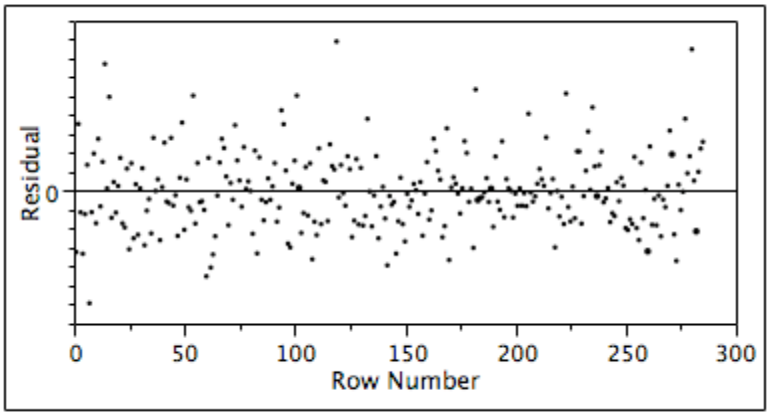
Analysis of Variance

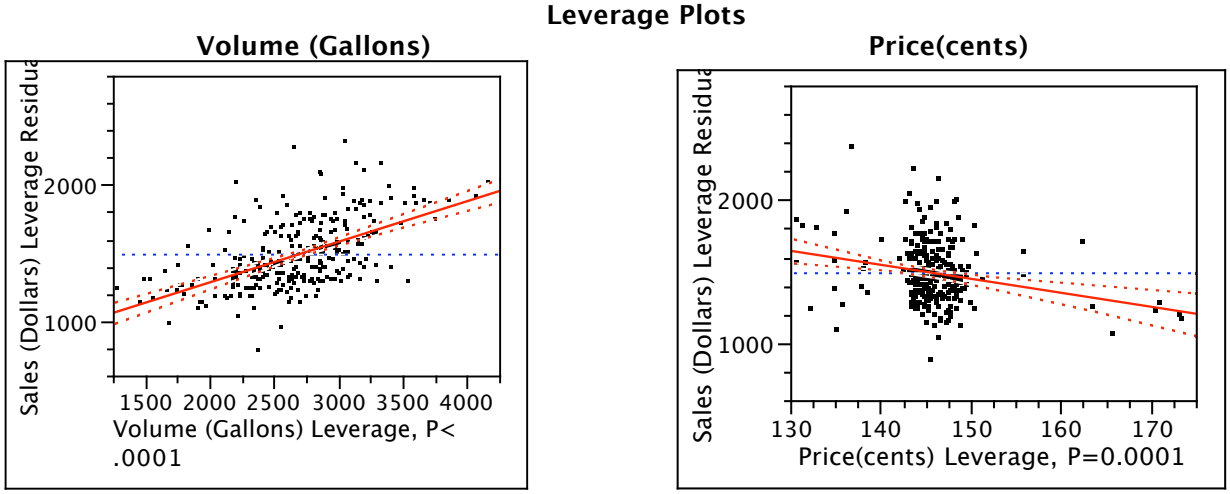
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	9688933	1937787	44.5191
Error	279	12144060	43527	**Prob > F**
C. Total	284	21832993		<.0001

Expanded Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2249.285	366.814	6.13	<.0001
Weekend[No]	-779.292	366.814	-2.12	0.0345
Weekend[Yes]	779.292	366.814	2.12	0.0345
Price(cents)	-9.957	2.795	-3.56	0.0004
Weekend[No]*Price(cents)	4.371	2.795	1.56	0.1191
Weekend[Yes]*Price(cents)	-4.371	2.795	-1.56	0.1191
Volume (Gallons)	0.301	0.047	6.41	<.0001
Weekend[No]*Volume (Gallons)	-0.050	0.047	-1.06	0.3445
Weekend[Yes]*Volume (Gallons)	0.050	0.047	1.06	0.3445







(32) Management would like to use this model to predict store sales at this location, assuming known values of the explanatory variables. Can management expect predictions from this model to come within 10% of the actual store sales during typical business periods?

(a) No, the value of R^2 is less than 0.90.
(b) No, the residuals from this model are not normally distributed.
(c) Yes, but only about 50% of the time.
(d) Yes, but only about 80% of the time.
(e) Yes, the fitted model explains statistically significant variation in the response.

(33) The fit of this model implies that, on average, the sale of another 20 gallons of gas on a weekday with the price set to $1.60 per gallon (typical a couple of years ago) results in

(a) $1 less in sales at the convenience store, on average.
(b) $1 more in sales at the convenience store, on average.
(c) $5 more in sales at the convenience store, on average.
(d) $6 more in sales at the convenience store, on average.
(e) $7 more in sales at the convenience store, on average.

(34) Based on the fit of this model, average sales at the convenience store on a weekend day on which the station sells 3,000 gallons priced at 160 cents per gallon ($1.60/gallon) is
(a) $1329
(b) $1559
(c) $1789
(d) $2338
(e) $2479

(35) On average, are sales at the convenience store higher on the weekend (Saturday or Sunday) or higher on a weekday?
(a) On a weekday, by a statistically significant margin.
(b) On a weekday, but the difference is not statistically significant.
(c) On the weekend, by a statistically significant margin.
(d) On the weekend, but the difference is not statistically significant.
(e) The shown analysis does not provide the information needed to answer this question.

(36) Regarding the effect of the price of gasoline on sales, the fit of this model indicates that, on average,
(a) Increasing prices reduce the volume of gasoline sold at the store.
(b) Increasing prices reduce the number of customers who shop at the store.
(c) Customers are more price-sensitive in their shopping during weekends.
(d) Customers buy more at the store on days when the price of gas is higher.
(e) The price of gas does not affect the sales of gas on weekends.

(37) The leverage plot for *Price* indicates that
(a) The price of gas is typically more than $1.50 at this location.
(b) The effect of Price in this model is not statistically significant.
(c) Collinearity has produced a large standard error for the estimated slope.
(d) The estimated slope depends heavily on the effects of leveraged outliers.
(e) The fitted model requires a transformation using logs to capture curvature.

(38) If the explanatory variable that measures the price of gasoline were expressed in dollars rather than cents (1 dollar = 100 cents), then
(a) The RMSE of the model would be 100 times smaller.
(b) The estimated slope for *Price* would be 100 times larger.
(c) The *t*-statistic for *Price* would be 100 times larger.
(d) The standard error for *Price* would be 100 times smaller.
(e) The *p*-value for *Price* would be 100 times smaller.

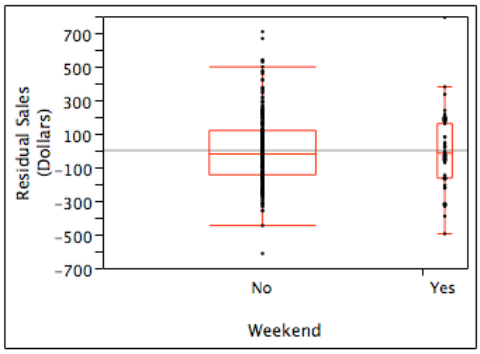
(39) If the interaction between *Price* and *Weekend* were removed from the model, then we can be sure that
(a) The R^2 of the model would increase, but not by a statistically significant amount.
(b) The R^2 of the model would decrease by a statistically significant amount.
(c) The R^2 of the model would decrease, but not by a statistically significant amount.
(d) The estimated slope for *Price* would increase.
(e) The estimated slope for *Price* would decrease.

(40) Suspecting fraud at this store, management used this model to estimate the total sales during the next 5 days, Monday through Friday. Given usual assumptions, these results suggest that the estimate of total sales for these 5 days should be accurate to (with the usual assumptions)
(a) ± $209, with 95% probability.
(b) ± $418, with 95% probability.
(c) ± $935, with 95% probability.
(d) ± $1045, with 95% probability.
(e) ± $2090, with 95% probability.

(41) If the explanatory variable *Car Wash* is added to the shown multiple regression, then from this output we can tell that
(a) The R^2 of the model would increase by 0.2085.
(b) The R^2 of the model would increase by 0.04347.
(c) The R^2 of the model would remain 0.444.
(d) The R^2 of the model would decrease.
(e) We cannot tell what would happen to R^2 from the shown information.

(42) An analyst split these data into subsets, one for weekdays and one for the weekends. Within the subset of weekdays, the analyst fit a multiple regression of sales (the response in this model) on *Price* and *Volume*. The slope for *Price* in that regression is
(a) –5.586
(b) –9.957
(c) –14.328
(d) –14.92
(e) Cannot be determined from the output given.

(43) The following plot suggests that the fitted model



(a) Violates the assumption of independence.
(b) Predicts sales more accurately for days on a weekend than during the week.
(c) Predicts sales more accurately for days during the week than the weekend.
(d) Violates the assumption of normality.
(e) Is OK from this perspective.

(44) The value of the Durbin-Watson statistic for this multiple regression would be approximately
(a) 0
(b) 1
(c) 2
(d) 3
(e) 4